الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne démocratique et populaire

وزارة التعليم السعالي والبحث العلمي

Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة

Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا

Faculté de Technologie

قسم الإلكترونيك

Département d'Électronique

Mémoire de Master

Filière : Télécommunications

Spécialité : Systèmes des télécommunications

Présenté par :

BOUSSELSELA Abderrahamne

IDENTIFICATION AUTOMATIQUE DU LOCUTEUR PAR LES RESEAUX DE NEURONNES RECURRENTS

Encadré par : BOUTALEB Nassima

Année Universitaire : 2024-2025

Remerciements

Au terme de ce travail, nous tenons à remercier en premier lieu Dieu (Allah) qui nous a donné la force, la volonté et le courage pour terminer cette mémoire

Dieu merci !

Toute notre gratitude et nos vifs remerciements vont à notre promoteur Dr. N.BOUTALEB Enseignant à l'Université de Blida pour avoir assuré l'encadrement de ce travail. Pour leurs aides, ses conseils et son suivi durant la réalisation de notre projet.

Un grand remerciement aux membres du jury chacun par son propre nom pour l'honneur et l'intérêt qu'ils nous ont accordé en acceptant d'examiner et d'évaluer notre mémoire.

Nos remerciements Vont particulièrement à nos parents et à toute notre famille pour leurs soutiens et amour inconditionnels et leurs encouragements durant tout notre parcours.

.

DEDICACES

Je dédie ce travail Aux deux personnes les plus nobles, précieux et les plus chères au Monde Ma mère, Mon père que DIEU les gardes.

A mon cher père qui n'a jamais cessé de m'encourager et de me Donner les conseils fructueux, qui a fait de son mieux pour assurer La continuité de mes études.

A ma très chère mère, Mère exemplaire pour mes frères et pour Moi même, tu as su donner l'éducation qu'il nous faut pour Affronter les épreuves de la vie.

Tu nous as comblés de ton amour maternel et tu répondais Présente à chacune de nos sollicitations.

Puisse le Tout Puissant T'accorder longue vie afin de profiter des fruits de ce labeur.

A mes soeurs et mes frères, pour leurs aides, disponibilités et précieux conseils, Que la vie vous apporte toute la joie et le bonheur. A mon binôme, Je tiens à exprimer ma profonde gratitude. Notre collaboration, notre complémentarité et notre travail d'équipe ont été la clé de notre réussite

LISTES DES ACRONYMES

CNN : Convolutional Neural Network

CPU : Central Processing Unit

GMM : Gaussian Mixture Model

FFT : Fast Fourier Transform

IA : Intelligence Artificielle

ID : Identification

GPU : Graphical Processing Unit

RAL : Reconaissance Automatique de Locuteur

RAP : Reconnaissance Automatique De la Parole

RNA : Réseau Neurone Artificiel

RNN : Recurrent Neural Network

RELU : Rectified Linear Unit

TD : Text Dependent

TI : Texte Independent

VAL : Vérification Automatique du Locuteur

ملخص :

يُعدّ التعرّف التلقائي على المتحدث، أو القياسات الحيوية للصوت، مجالًا سريع التطور بفضل التطورات في التفاعل بين الإنسان والحاسوب .في هذا العمل، نستكشف نهجًا لتحديد المتحدث قائمًا على التعلّم العميق من الصوت .تُستخدم معاملات التردد السيبستري (MFCC)لتمثيل السمات الصوتية .نستغل الشبكات العصبية المتكررة (RNNs)، ومتغيراتها LSTMوGRU، المُكيّفة لمعالجة التسلسلات الزمنية .تُمكّننا هذه البنى من نمذجة ديناميكيات الصوت بشكل أفضل .وأخيرًا، نقارن أداءها لتحديد أكثر الأنظمة كفاءة.

كلمات المفاتيح:

التعرف الآلي على المتحدث، المعاملاتMFCC، شبكات الخلايا العصبية المتكررة RNN ، التعليم المعمقLSTM,GRU.

Résumé:

La reconnaissance automatique du locuteur, ou biométrie vocale, est un domaine en pleine évolution grâce aux avancées de l'interaction homme-machine. Dans ce travail, nous explorons une approche d'identification du locuteur basée sur l'apprentissage profond à partir de la voix. Les coefficients cepstraux en fréquences de Mel (MFCC) sont utilisés pour représenter les caractéristiques acoustiques. Nous exploitons les réseaux de neurones récurrents (RNN), et leurs variantes LSTM et GRU, adaptés au traitement des séquences temporelles. Ces architectures permettent de mieux modéliser la dynamique vocale. Enfin, nous comparons leurs performances pour identifier les systèmes les plus efficaces

.**Mots clés :**

Reconnaissance automatique du locuteur, MFCC, RNN, apprentissage profond, LSTM, GRU

Abstract :

Automatic speaker recognition, or voice biometrics, is a rapidly evolving field thanks to advances in human-computer interaction. In this work, we explore a speaker identification approach based on deep learning from voice. Mel frequency cepstral coefficients (MFCC) are used to represent acoustic features. We exploit recurrent neural networks (RNNs), and their variants LSTM and GRU, adapted to the processing of temporal sequences. These architectures allow us to better model voice dynamics. Finally, we compare their performances to identify the most efficient systems.**Keywords:**

Automatic Speaker Recognition, MFCC, RNN, deep learning, LSTM, GRU

TABLES DES MATIERES

LIST DES TABLAUE

LIST DES FIGURE

INTRODUCTION GENERAL

Le domaine de l'identification automatique du locuteur est en plein essor au croisement de la linguistique, de la phonétique et de l'intelligence artificielle et a pour but de vérifier, en partant d'un échantillon vocal, si une voix non identifiée peut être rattachée à une personne donnée sur la base de traits distinctifs de la parole humaine ; cette technologie est mise en œuvre dans des cas où sa démonstration revêt un caractère délicat tels que les enquêtes pénales, la sécurisation de systèmes de traitement de l'information ou les services sur mesure.-

Ce travail de recherche vise à réaliser un système d'identification automatique du locuteur basé sur l'apprentissage profond et les réseaux de neurones récurrents (RNN) en combinant la reconnaissance des voix avec l'utilisation de coefficients cepstraux en fréquence Mel (MFCC) pour ce faire dans la présentation des résultats de recherche étalonnés par des systèmes standards.

Pour ce qui est du cadre de mémoire choisi, celui-ci s'organise en trois parties.

La première de celles-ci est une revue de littérature consacrée à l'identification automatique du locuteur, aux approches qui ont été utilisées et aux différentes possibilités d'applications.

Dans le deuxième chapitre, les techniques d'apprentissage profond et en particulier les réseaux de neurones récurrents sont abordés, ainsi que les méthodes d'extraction de caractéristiques dans la reconnaissance vocale.

Le dernier chapitre présente la méthodologie proposée pour construire le système d'identification automatique du locuteur, les expériences effectuées ainsi que les résultats obtenus. Cette étude vise à aller vers une amélioration de la reconnaissance vocale ou de l'identification de locuteurs à partir des potentialités offertes par des systèmes basés sur l'apprentissage profond afin de mettre au point un système capable de retrouver les locuteurs de façon efficace.

CHAPITRE I : IDENTIFICATION AUTOMATIQUE DU LOCUTEUR

CHAPITRE I : IDENTIFICATION AUTOMATIQUE DU LOCUTEUR

Introduction

L'homme utilise la parole, un outil de communication extrêmement efficace et instinctif. Il nourrit depuis longtemps le rêve de communiquer avec des machines par ce même canal, ce qui accroîtrait leur intelligence.

La recherche sur identification automatique du locuteur, qui se trouve à l'intersection de la linguistique, de la phonétique et de l'intelligence artificielle, connaît une expansion considérable. L'objectif est de vérifier, à partir d'un échantillon vocal, si une voix non identifiée peut être associée à une personne spécifique, en s'appuyant sur des traits distinctifs de la parole humaine Cette technologie est de plus en plus mise en œuvre dans des situations délicates telles que les investigations criminelles, la protection des systèmes d'information ou les services sur mesure [9][13]

L'analyse précise du signal vocal, contenant non seulement des informations linguistiques mais également des indices biométriques uniques à chaque individu, est au cœur du traitement automatique de la parole Toutefois, ce signal présente une variabilité importante qui complique son analyse. À cette fin, diverses méthodes - phonétique, acoustique et automatique - sont employées pour déterminer, mettre en parallèle et analyser les traits vocaux dans le but de confirmer ou d'identifier un intervenant [12]

I.1. Identification du locuteur en forensique

L'identification du locuteur en contexte forensique vise à déterminer si un enregistrement vocal inconnu peut être attribué, avec un certain degré de certitude, à une personne suspectée. Cette discipline, à l'intersection de la linguistique, de la phonétique et de l'ingénierie, joue un rôle de plus en plus important dans les enquêtes judiciaires, notamment dans les affaires de menaces téléphoniques, de fraudes ou de terrorisme [3]

L'analyse forensique de la parole repose sur le postulat que chaque individu possède une voix unique, résultant de caractéristiques anatomiques (telles que la forme du conduit vocal) et comportementales (comme le rythme, l'intonation ou l'accent). L'objectif est de comparer un

échantillon vocal de référence (provenant du suspect) à un échantillon inconnu (issu d'un enregistrement) pour évaluer la probabilité qu'ils aient été produits par le même locuteur

On distingue généralement trois deux approches (tableau I.1):

Table 4.1: Comparaison des approches d'indentification du locuteur.

Approche	Description	Avantages	Inconvénients
Phonétique	Analyse auditive et/ou visuelle (spectrogrammes) réalisée par des experts humains.	Utile en cas d'enregistrements de mauvaise qualité ; permet une approche qualitative.	Subjective ; dépend de l'expérience de l'expert ; peu reproductible.
Acoustique-instrumentale	Extraction de mesures acoustiques (formants, F0, durée, etc.) et comparaison statistique via logiciels spécialisés.	Fournit des données objectives ; complète l'analyse perceptive.	Sensible aux variations de canal et d'enregistrement ; nécessite un bon contrôle des variables.
Automatique	Utilisation de caractéristiques comme les MFCC et d'algorithmes (GMM, i-vectors, x-vectors) pour la reconnaissance.	Rapide, reproductible, efficace sur de grandes bases de données ; performante avec les systèmes récents.	Nécessite beaucoup de données et des ressources de calcul ; résultats sensibles aux données d'entraînement.

I.2 Processus de la Parole

La langue est l'un des outils de communication les plus élaborés et sophistiqués chez l'homme. Elle découle de la coordination minutieuse entre les organes respiratoires, phonatoires et articulatoires pour générer des sons significatifs. La parole, en tant que signal sonore, ne transmet pas uniquement le contenu linguistique, elle véhicule également des informations relatives à l'identité, à l'émotion et à l'intention de celui qui parle.

Dans le domaine du traitement du signal, la parole joue un rôle central grâce à ses propriétés acoustiques distinctives, qui proviennent des mécanismes particuliers de sa production.

I.2.1. Production de la parole

Différents organes participent à la production de la parole. Le flux d'air provenant des poumons est à l'origine de la production de la voix.

Ce courant d'air va passer par le larynx pour provoquer ou pas la vibration des cordes vocales. Il traversera ensuite le passage vocal (cavité nasale et buccale) ainsi que les éléments articulatoires tels que les lèvres et la langue (Figure I.1). Ce système fonctionne comme un filtre, perçu comme linéaire, dont la réponse impulsionnelle contient des fréquences de résonance identifiées par des crêtes, nommées formants, dans le spectre du signal sortant.

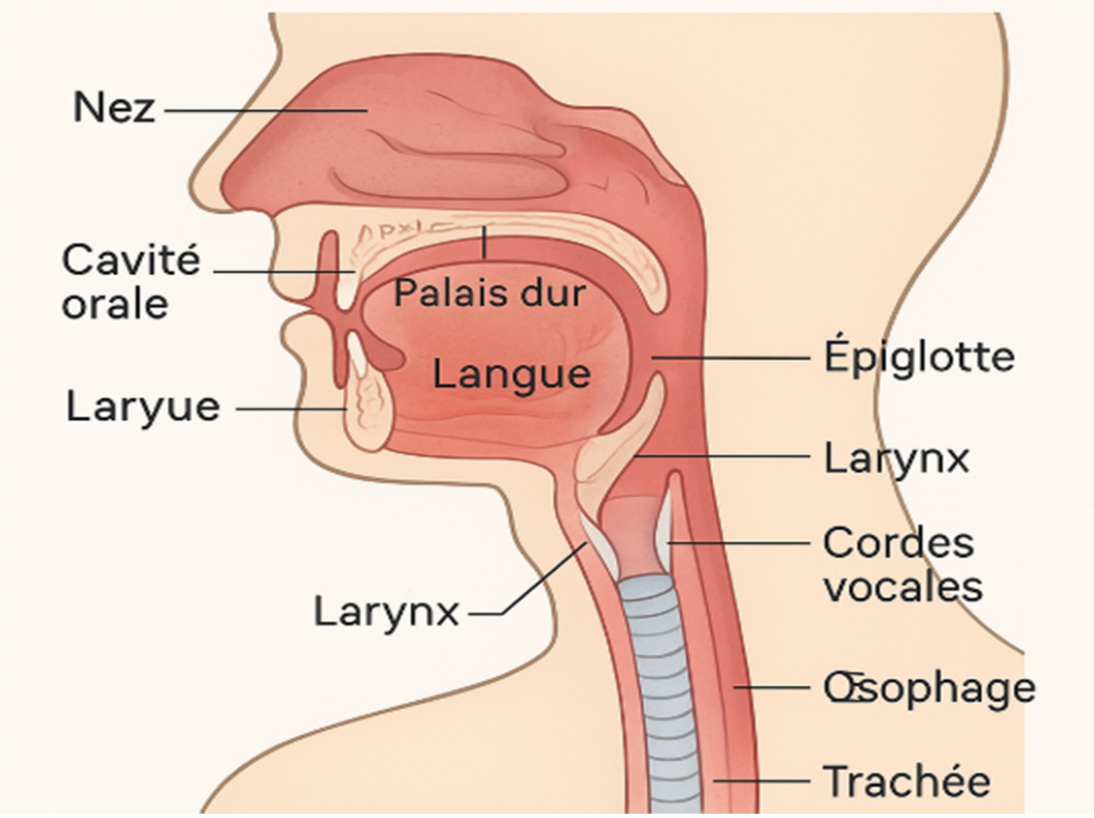


Figure 3.1: *Modèle physiologique de la production de la parole.*

Le signal obtenu est généralement non stationnaire, mais peut être perçu comme stationnaire sur de brèves durées, autour de 20ms (signal pseudo-stationnaire). Pour un segment de discours de cette durée, la voix est généralement et schématiquement divisée en deux catégories distinctes :

1.Voisée : lorsqu'il y a vibration des cordes vocales, le son est alors presque périodique et on parle de voisement.

2. **Non voisée:** Dans le cas d'un simple souffle, le signal est alors jugé comme étant aléatoire s'il n'est pas visible.

I.2.2. Dynamiques du signal de parole

Le signal de parole est soumis à une grande variabilité, ce qui rend son traitement et son analyse particulièrement complexes. Cette variabilité peut être attribuée à plusieurs facteurs, notamment les différences inter-locuteurs, intra-locuteurs, linguistiques, contextuelles et environnementales donnes un referance de laparagraphe

I.2.2.1. Dynamique inter-locuteur

La variabilité inter-locuteur reflète les différences vocales entre individus, dues à des facteurs anatomiques, physiologiques et sociolinguistiques. Des éléments comme le larynx, les cavités vocales, le sexe et l'âge influencent la voix. Par exemple, les hommes ont généralement une voix plus grave. Ces différences compliquent le traitement automatique de la parole. Adapter les systèmes à cette diversité constitue un enjeu majeur en reconnaissance vocale. Fant, G. (1960)

I.2.2.2. Dynamique intra-locuteur

Correspond aux changements dans la parole d'un même individu selon le contexte. Elle peut être influencée par l'émotion, la fatigue, la santé, ou le style de communication. Ces variations affectent le débit, l'intonation et la clarté de la parole. Elles compliquent la reconnaissance automatique, car le signal n'est pas stable. Cette variabilité représente un défi pour les systèmes de traitement vocal. Laver, J. (1994)

I.3. Reconnaissance vocale

En règle générale, la reconnaissance fait référence à l'aptitude à identifier, comprendre ou différencier quelque chose. Toutefois, cette notion peut être déployée dans différents secteurs, où elle acquiert des sens et des conséquences particuliers. La reconnaissance offre diverses applications, notamment dans le secteur de la reconnaissance vocale, dont les suivantes peuvent être identifiées :

- Reconnaissance de la parole.
- Reconnaissance de locuteur
- Reconnaissance vocale pour la transcription.

Notre projet met l'accent principalement sur les deux premiers domaines.

I.3.1. Système de reconnaissance de la parole

La reconnaissance automatique de la parole soulève plusieurs enjeux d'ordre théorique. Du fait de leurs complexités, seul un certain nombre de sous-problèmes a pu être résolu jusqu'à présent. Ces solutions incomplètes sont associées à des contraintes de différents niveaux, et les systèmes en place présupposent une collaboration variable de la part des utilisateurs.

Pour classer les systèmes de reconnaissance automatique, on a généralement recours aux critères suivants :

- L'articulation des mots : des syllabes ou mots séparés jusqu'à des mots liés, allant jusqu'à une parole dite « continue », c'est-à-dire sans interruptions artificielles.
- Dimension du vocabulaire et complexité de la grammaire (le degré de sophistication du langage permis).
- Le degré variable de dépendance à l'égard du locuteur.
- La protection de l'environnement (résistance aux conditions d'enregistrement).

De plus, il serait pertinent de les distinguer en fonction de deux critères qui sont également significatifs :

- Est-il nécessaire de comprendre ou pas ? (Un système de compréhension tente de saisir le sens de l'énoncé oral.)
- Le discours est-il naturel, ou la syntaxe des phrases doit-elle être contraignante ?

Les systèmes élaborés grâce à la RAP sont prévus pour des applications particulières. Ceci entraîne une limitation de l'étendue du dialogue entre l'homme et la machine.

Le but primordial des recherches sur la reconnaissance et la compréhension automatiques de la parole est « d'aspirer, à long terme, à une interaction aussi naturelle que possible entre l'utilisateur et le dispositif, dans le contexte d'une application spécifique »[3].

I.3.1.1. Domaines d'application de la RAP

Il existe une multitude d'applications pour la reconnaissance vocale. Elle permet une utilisation totalement libérée des mains et de la vue, laissant l'utilisateur maître de ses gestes. Dans la RAP, la transmission d'informations est plus rapide que celle offerte par l'utilisation d'un clavier. Finalement, presque tout le monde sait communiquer verbalement, bien que rares soient ceux qui sont totalement exemptés des erreurs de frappe, d'orthographe, etc.

Ces bénéfices sont si significatifs qu'on trouve déjà sur le marché des appareils à usage restreint, mais pourtant performants.

- Citons certaines applications qui ont déjà vu le jour :
- Saisie vocale de données.
- Émet des directives pendant la conduite d'une voiture ou d'un avion.
- Aide aux handicapés.
- Salle d'hôpital équipée de commandes vocales pour le patient.
- Utilisation de la commande vocale pour contrôler des machines ou des robots.
- Activation vocale d'une montre portable, entre autres

I.3.1.2. Analyse de la complexité de la parole

Le signal de la parole est extrêmement complexe et soumis à de nombreuses variations. Étant donné qu'il a été généré par un système phonatoire humain sophistiqué, le caractériser uniquement à partir d'une représentation bidimensionnelle de diffusion des ondes s'avère complexe. On peut identifier plusieurs de ses attributs tels que les phonèmes ou sons fondamentaux, la hauteur tonale, le timbre acoustique, l'amplitude sonore, la rapidité... En réalité, la voix est bien plus complexe que ce qu'on peut percevoir simplement par l'oreille. Non seulement l'onde sonore change en fonction des sons prononcés, mais elle varie aussi selon les intervenants. L'immense diversité que peut offrir une même allocution en fonction de la manière dont le locuteur s'exprime, susceptible de chanter, de créer, de murmurer, d'être enroué ou enrhumé, mais aussi en tenant compte du locuteur lui-même (homme, femme, enfant, voix nasillarde, différences de timbre), Sans tenir compte des accents régionaux, la définition des invariants devient particulièrement complexe. Il est nécessaire de distinguer ce qui définit les phonèmes, qui devraient rester constants indépendamment du locuteur et de sa manière de prononcer, de l'aspect spécifique à chaque interlocuteur. Comment notre cerveau parvient-il à différencier un mot d'un autre, indépendamment de l'identité de la personne qui parle ?

Par ailleurs, l'évaluation du signal vocal est grandement affectée par la fonction de transfert du système de reconnaissance (les dispositifs d'acquisition et de transmission) et par l'environnement. Le principal défi pour obtenir une précision élevée dans la reconnaissance vocale réside dans la grande variabilité des caractéristiques d'un signal vocal. La complexité du signal de parole découle de l'interaction de plusieurs éléments : la redondance inhérente au signal acoustique, la variation significative entre différents orateurs et interlocuteurs, les conséquences de la coarticulation dans une conversation ininterrompue, ainsi que les conditions d'enregistrement [4].

I.3.1.3. Structure du signal vocal

La parole est un phénomène naturel, évolutif dans le temps, qui peut être directement illustré sous la forme d'un signal analogique. C'est un vecteur acoustique qui transporte des informations de grande complexité,

variabilité et redondance.

L'analyse de ce type de signal est complexe en raison du grand nombre de paramètres qui lui sont liés. Toutefois, trois éléments essentiels se distinguent : la fréquence fondamentale, le spectre fréquentiel et l'énergie. On appelle traits acoustiques ces paramètres, et ils sont listés ci-dessous [5, 6].

Fréquence fondamentale (F0)

C'est une propriété acoustique spécifique à chaque individu. Elle dépend de divers facteurs physiologiques, comme la taille de la glotte et la longueur de la trachée. Elle est caractérisée par le rythme du cycle d'ouverture et de fermeture des cordes.

Vocales lors de la production des sons voisés. La fréquence fondamentale varie d'un orateur à l'autre en fonction du sexe et de l'âge de la figure suivante :

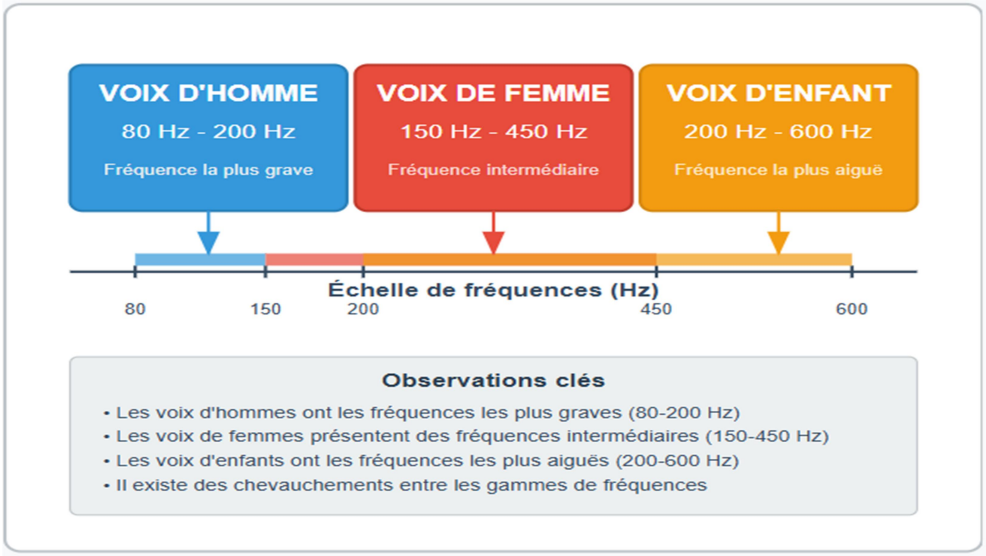


Figure 1.2 : *Les gammes de frequence de la voix humaine.*

Spectre de fréquence

Il s'agit de la représentation d'un signal dans le domaine des fréquences (ensemble de fréquences progressant selon une progression arithmétique). Un trait distinctif essentiel pour reconnaître chaque intervenant par sa voix, connu sous le nom de timbre.

Energie

Elle se réfère à l'intensité sonore. Elle est généralement plus forte pour les segments de parole qui sont voisés que pour ceux qui ne le sont pas.

Spectrogramme

Le spectrogramme est une représentation graphique associant à chaque moment d'un signal, son spectre fréquentiel. On se sert des spectrogrammes pour reconnaître des bruits, tels que les cris d'animaux et les sons issus d'instruments de musique. On les utilise couramment dans le secteur de la reconnaissance vocale.

On pourrait définir un spectrogramme comme une représentation qui montre l'intensité d'un signal dans le temps à différentes fréquences d'une onde. Les spectrogrammes peuvent se présenter sous forme de diagrammes en deux dimensions, où une troisième variable est indiquée par des couleurs, ou bien sous forme de courbes en trois dimensions intégrant une quatrième variable colorimétrique.

Amplitude

L'amplitude représente une évaluation de la variation d'un son sur une unique période (comme le temps ou la période spatiale). L'amplitude d'un signal non périodique correspond à son amplitude par rapport à une valeur de référence. On retrouve plusieurs définitions de l'Amplitude, toutes basées sur l'écart entre les valeurs limites de la variable. Dans les documents plus anciens, la phase d'une fonction périodique est parfois désignée sous le terme d'amplitude [8].

I.3.2. Reconnaissance automatique du locuteur RAL

Les systèmes de RAL sont largement utilisés dans les domaines de la sécurité, de l'authentification à distance, et des services personnalisés. Ils peuvent être **texte-dépendants**, lorsqu'une phrase spécifique est requise, ou **texte-indépendants**, lorsque n'importe quel contenu vocal peut être utilisé pour l'identification [32]

Une première méthode pour classer les systèmes de reconnaissance vocale est de les organiser en fonction des tâches qu'ils doivent réaliser.

I.3.2.1. Vérification automatique du locuteur

Cela vise à confirmer que la personne qui s'exprime est bien celle attendue ou déclarée. Divers termes présents dans la littérature pourraient correspondre à la même définition que la vérification du locuteur (VL), tels que la vérification de voix, l'authentification vocale et l'authentification du locuteur.[19]

Elle réalise une comparaison (également appelée décision binaire) entre les caractéristiques de la voix d'entrée et celles des voix proclamées qui sont stockées dans une base de données.

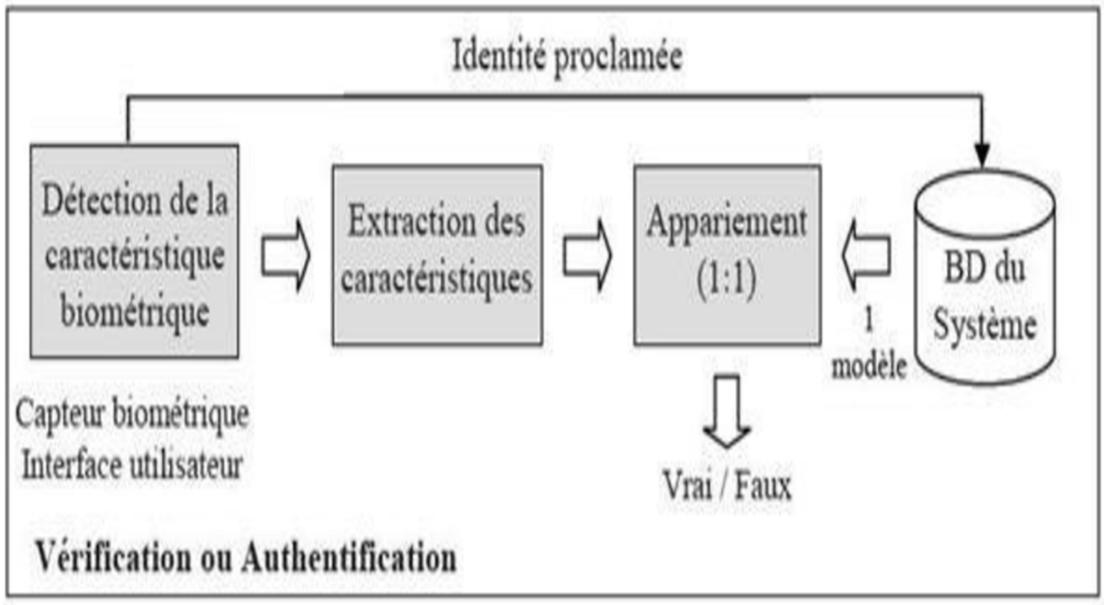


Figure 4.3: *Système de vérification du locuteur. [12]*

On utilise le traitement préalable du signal d'entrée pour formater le signal afin de dégager les paramètres significatifs. On appelle également cette phase analyse acoustique.

Suite à cette procédure, nous avons des vecteurs acoustiques. Ces vecteurs acoustiques servent à obtenir un modèle spécifique pour chaque intervenant, qui est ensuite stocké dans une base de données de référence. L'étape de vérification consiste à comparer le modèle du locuteur revendiqué, stocké dans la base de données, avec le modèle du signal vocal qui est introduit dans le système. Si le score déterminé dépasse un certain seuil, l'identité déclarée est vérifiée. Si l'on fixe un critère strict, le système assure une haute sécurité et empêche l'admission de fraudeurs. Cependant, cela pourrait également entraîner le refus d'identités authentiques, et inversement.[11]

I.3.2.2. Identification automatique du locuteur

L'apprentissage est similaire pour l'identification et la vérification, mais la reconnaissance diffère : comparaison parallèle des modèles pour trouver le plus probable dans l'identification

On distingue deux formes d'identification du locuteur : l'identification dans un ensemble restreint, où le locuteur est reconnu parmi un groupe spécifique de L locuteurs, et l'identification dans un cadre élargi, où le locuteur pourrait ne pas appartenir au groupe de référence. Dans ce dernier scénario, on peut envisager le processus comme une identification en groupe restreint suivie d'une validation de l'identité suggérée.

On peut également classer les systèmes de reconnaissance vocale selon leur dépendance au contenu verbal : les systèmes indépendants du discours n'ont pas besoin d'un contenu linguistique spécifique, alors que les systèmes dépendants du discours nécessitent que l'utilisateur prononce un texte spécifique pour effectuer la reconnaissance.

Le fonctionnement d'un système de reconnaissance du locuteur se décompose en deux phases : une phase d'apprentissage et une phase de test.

1. ***Phase d'apprentissage***

Lors de la phase d'apprentissage, un locuteur prononce un ensemble de mots ou de phrases pour créer un dictionnaire de références acoustiques dans la machine. Selon l'approche analytique, l'utilisateur est invité à énoncer des phrases spécifiques, souvent sans signification particulière, mais conçues pour inclure des combinaisons de sons (phonèmes) représentatives.

2. ***Phase de test***

En phase de test, les systèmes d'identification et de vérification fonctionnent différemment. L'identification consiste à identifier un locuteur inconnu à partir d'un énoncé, tandis que la vérification consiste à valider ou invalider une identité revendiquée en fonction d'un énoncé de test

I.3.2.3. Application des systèmes RAL

Les applications de l'IAL (identification automatique du locuteur) sont diverses et peuvent différer en fonction de leurs catégories. Les progrès dans les méthodes de l'ASR ont permis aux systèmes de progresser et d'accroître progressivement leur efficacité. De plus, la majorité des systèmes ASR

s'agit de systèmes qui peuvent être dépendants ou indépendants du locuteur. Les systèmes basés sur le locuteur nécessitent une phase d'entraînement durant laquelle un grand nombre d'heures de discours est souvent requis. Toutefois, les systèmes non dépendants du locuteur n'ont pas besoin d'une phase d'apprentissage des données et sont préférables pour plusieurs applications où l'apprentissage est compliqué à réaliser. Cette partie présente succinctement les quatre principales catégories de systèmes existants dans le domaine de la reconnaissance vocale. [31]

a. Instructions vocales

L'identification vocale automatique sert à authentifier un individu en se basant sur sa voix, fonctionnant comme une clé biométrique. Elle est employée dans des secteurs délicats tels que la

finance ou les résidences intelligentes. Cette approche améliore la sécurité tout en demeurant naturelle et sans contact.

b. Personnalisation dans les aides vocales.

La technologie de reconnaissance vocale donne la possibilité aux assistants virtuels de reconnaître les utilisateurs et d'adapter leurs réponses en fonction d'eux. Elle personnalise des services tels que le calendrier, la musique ou les notifications en fonction des goûts individuels. Ceci optimise l'expérience de l'utilisateur et intensifie l'interaction entre l'homme et la machine. Des assistants tels que Google Assistant ou Alexa emploient cette fonctionnalité pour administrer divers profils vocaux.

c. Transcription et séparation des locuteurs.

La transcription permet de transformer la parole en écriture, alors que la diarisation sert à distinguer les divers intervenants dans un enregistrement audio. Ces éléments combinés permettent de générer des documents précis indiquant qui a déclaré quoi. Ces outils sont indispensables pour les conférences, les entretiens et les discussions.

Conclusion

Dans cette section, nous avons mis en avant la reconnaissance du locuteur comme l'un des travaux de pointe de l'intelligence artificielle. Elle vise à reproduire la compétence humaine consistant à extraire des informations à partir de la parole d'autrui. Cette mission, trop délicate pour qu'un seul système informatique puisse la réaliser, a été décomposée en divers sous-tâches selon la nature des informations à identifier et à reconnaître. Les questions les plus explorées incluent l'identification du locuteur, la détection de son état d'âme, la détermination de la langue utilisée et l'analyse du langage parlé. Généralement, on aborde le problème de la RAL sous deux perspectives : analytique et globale.L'approche analytique facilite la résolution du problème de la reconnaissance continue du locuteur, tandis que l'approche globale, tirée des méthodes de reconnaissance de formes, met l'accent sur l'aspect acoustique plutôt que sur l'aspect linguistique. Cette méthode constitue une solution pour esquiver les problématiques liées à l'analyse linguistique. L'issue du système est déterminée suite à l'évaluation d'un coefficient de similarité entre la forme à identifier et un ensemble de formes déjà enregistrées. Chaque méthode possède ses propres atouts et faiblesses, cependant les méthodes analytiques se révèlent plus performantes que les méthodes globales en termes de qualité des résultats, surtout lorsque le volume d'énoncés possibles est important et/ou lorsque la redondance acoustique est faible. Chaque année, ils s'améliorent de plus en plus.Qu'elle soit globale ou analytique, l'identification du locuteur débute par un traitement acoustique préalable du signal vocal

visant à minimiser le volume d'informations et à supprimer les redondances apparentes. En dépit de ces obstacles et défis, les systèmes de reconnaissance automatique de la parole s'améliorent constamment d'année en année.

CHAPITRE II : APPRENTISSAGE PROFOND RNN

CHAPITRE II : APPRENTISSAGE PROFOND RNN

Introduction

L'intelligence artificielle (IA) est une discipline de l'information qui vise a reconduire, a l'aide de machines, des capacités cognitives humaines telles que la perception, l'apprentissage, ou la prise de décision. Elle regroupe plusieurs sous-domaines, dans l'apprentissage automatique et l'apprentissage profond. Ces techniques aux systèmes de s'améliorer par l'expérience, sans être explicitement programmes pour chaque tache. Grace aux réseaux de neurones artificielle, l'IA connu des progrès spectaculaires dans des domaine comme la reconnaissance d'images, le traitement du langage et ;a robotique. L'apprentissage peut êtres supervise, non supervise ou par renforcement, selon le type de données disponibles et l'objectif vise. Inspirée du fonctionnement du cerveau humain, l'IA reposé sur des modèles.

Ce chapitre propose d'explorer les fondements, les types d'apprentissage, ainsi que les architectures neuronales a la base des systèmes intelligents. L'étude de ces concepts est essentielle pour comprendre les avancées actuelles et les perspectives de l'intelligence artificielle

II.1. Intelligence artificielle

L'intelligence artificielle (IA) est un domaine scientifique consacré à l'élaboration de programmes qualifiés d'intelligents. Ces logiciels sont élaborés pour traiter des problématiques qui, jusqu'à récemment, étaient vues comme des capacités purement humaines.

Il convient de noter que l'intelligence artificielle comprend l'apprentissage automatique, également appelé machine learning, qui englobe à son tour l'apprentissage profond ou deep learning. Ces trois notions sont profondément connectées et dépendantes les unes des autres, constituant un cadre cohérent pour l'élaboration de systèmes intelligents [12][16]

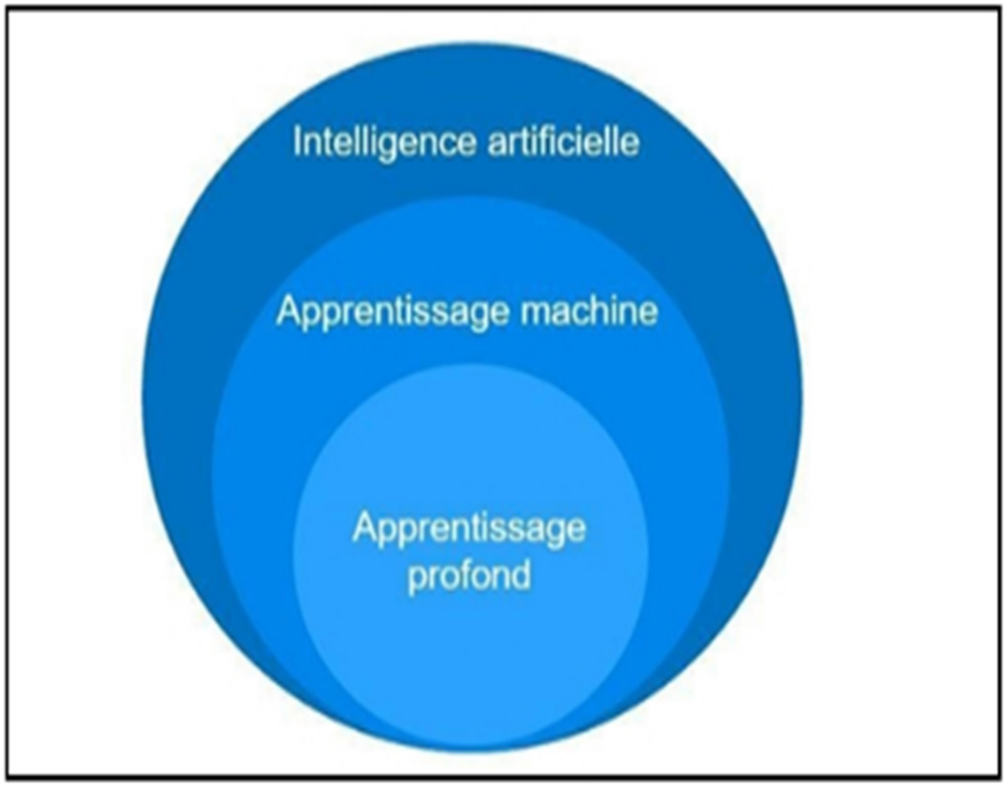


Figure 2.1 : *Les différentes zones de l'intelligence artificielle***.**

II.1.1. APPRENTISSAGE AUTOMATIQUE

Le machine learning, ou apprentissage automatique en français, est un sous-domaine de l'intelligence artificielle qui cherche à doter les machines de la faculté d'apprendre à partir de données sur la base de modèles mathématiques. L'idée fondamentale est de tirer des renseignements appropriés à partir d'un corpus de données d'apprentissage.

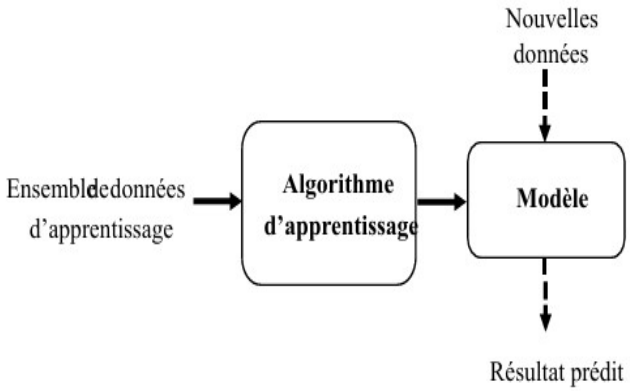


Figure 2.2 : Processus de l'apprentissage machine

L'objectif primordial de cette étape est de définir les paramètres optimaux d'un modèle pour garantir des performances maximales pendant l'accomplissement de la tâche qui lui est confiée. Après l'achèvement de la formation, le modèle peut être mis en production pour une application concrète [11].

L'apprentissage automatique se subdivise en différents types, chacun étant défini par la nature des tâches à accomplir. Dans les prochaines sections, nous examinerons les principaux types d'apprentissage. [16].

II.1.2. APPRENTISSAGE PROFOND

L'apprentissage profond, qui est une branche de l'apprentissage automatique, sert à former des systèmes informatiques connus sous le nom de réseaux de neurones artificiels (RNA). Ces méthodes reposent sur des algorithmes qui imitent le mode de fonctionnement du cerveau humain. Les réseaux de neurones artificiels ont la capacité de traiter des problématiques complexes comme l'identification d'images, l'interprétation de la voix et le traitement du langage naturel.

Les structures multicouches de neurones ou de nœuds artificiels reliés par diverses sortes de liaisons sont exploitées par les algorithmes d'apprentissage profond. Ces liaisons sont conçues pour reconnaître et saisir les traits présents dans un jeu de données particulier. Cette structure offre aux algorithmes la possibilité de se perfectionner à partir de leurs expériences et d'améliorer ainsi leur efficacité dans l'exécution des missions confiées [13].

II.2. TYPES D'APPRENTISSAGE

II.2.1. Apprentissage supervision

L'apprentissage supervisé se fonde sur l'emploi d'un jeu d'exemples labellisés, fournis par un superviseur qualifié ou un expert humain. Chaque exemple comporte une explication d'une situation, ainsi qu'une étiquette (une classe, symbolisée par des valeurs numériques ou nominales) qui définit l'action appropriée que le système se doit d'exécuter dans ce contexte.

Le but de l'apprentissage supervisé est d'habiliter le système à généraliser ses réponses pour pouvoir gérer correctement des situations nouvelles, qui ne figurent pas dans l'ensemble d'entraînement. Concrètement, l'utilisateur fournit à l'algorithme des paires d'entrées et de sorties attendues (X, y)(X, y), comme illustré dans la Figure II.3. L'algorithme doit donc apprendre à associer les entrées aux sorties, afin de générer une sortie appropriée pour des données qu'il n'a jamais rencontrées auparavant [3].

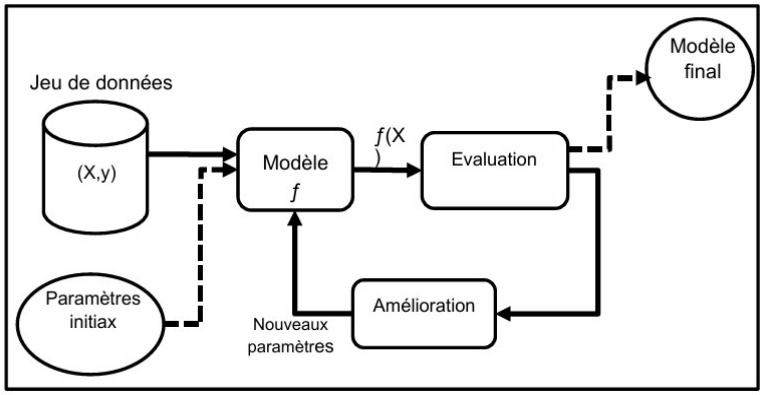


Figure 2.3 : Processus de l'apprentissage supervision [2]

On distingue deux grands types de problèmes d'apprentissage supervisé : la régression et la classification.

a) Classification : Dans le cadre de la classification, le jeu de données employé contient un nombre limité de catégories, et chaque instance est attribuée à l'une de ces catégories. Chaque exemple possède une cible qui est une valeur distincte correspondant à une catégorie précise. Un modèle d'apprentissage automatique est formé sur ces données pour assigner des classifications aux entrées.

Par exemple, pour la tâche de classification de documents, on pourrait envisager un jeu de données qui contient trois classes distinctes : « Économie », « Politique » et « Autre ». Le modèle est formé

pour examiner un document et décider s'il traite de l'économie, de la politique ou d'un autre thème, puis il associe le document à une catégorie symbolisée par une valeur précise [16].

b) Régression : Dans les cas de régression, la variable dépendante est composée d'une ou plusieurs valeurs continues. Un modèle de machine learning est formé pour anticiper ces valeurs réelles.

Un exemple typique de problème régressif est la prévision du temps, comme l'estimation de la température. Dans cette situation, la cible est une valeur continue. De façon comparable, des facteurs additionnels comme la pression atmosphérique ou le niveau d'humidité peuvent être intégrés dans la variable cible, créant ainsi un vecteur de valeurs continues à prédire [16].

II.2.2. Apprentissage sans supervision

L'apprentissage non supervisé se produit lorsque les exemples à disposition ne sont pas labellisés et qu'aucune détermination préalable n'est faite quant au nombre ou à la nature des classes. Ce genre d'apprentissage, aussi connu sous le nom de regroupement ou clustering, n'exige pas de compétence humaine préexistante. L'algorithme doit être capable de dévoiler, de façon autonome, la structure parfois dissimulée des données.

Dans ce contexte, le système examine les données selon leurs caractéristiques disponibles et les classe en groupes homogènes basés sur leurs similitudes. On évalue généralement ces similarités via une fonction de distance entre les exemples en paires. Après la formation des groupes, c'est à l'opérateur de donner une interprétation ou d'assigner un sens aux divers groupes identifiés [18].

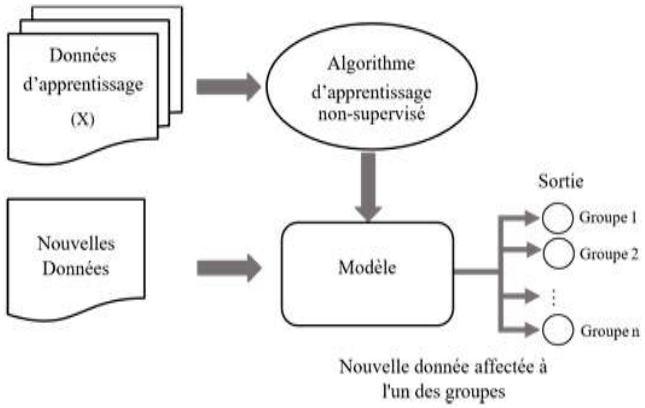


Figure 2.4 : *Processus d'apprentissage sans supervision [2].*

II.3. Problèmes abordés dans l'apprentissage sans supervision

a).Classification : le clustering, ou classification par regroupement, est une des méthodes les plus prisées et essentielles dans le cadre de l'apprentissage non supervisé. Cela implique d'analyser les données afin de repérer les groupes sous-jacents. L'algorithme identifie des schémas et organise les données selon leur ressemblance.

Cette technique permet de définir le nombre de groupes à détecter. On peut par la suite classer le regroupement en différentes catégories, y compris :

- Excusif : Chaque donnée est associée à un seul groupe.
- Agglomérat : L'élaboration des groupes se fait progressivement par l'ajout d'exemples.
- Superposition : Un exemple peut faire partie de plusieurs catégories.
- Probabiliste : Les informations sont attribuées à des catégories en fonction d'une probabilité.

b).Diminution : de la dimensionnalité : Les techniques de diminution de la dimensionnalité servent à simplifier les données en réduisant le nombre de variables (ou attributs) tout en conservant les informations clés.

En raison de l'abondance des caractéristiques, les algorithmes peuvent se révéler compliqués à interpréter ou à manipuler, certaines d'entre elles pouvant être redondantes ou hautement corrélées. Les techniques de réduction de dimensionnalité servent à diminuer la complexité des données en isolant ou en choisissant les attributs les plus significatifs.

On classe généralement ces techniques en deux types distincts :

Extraction de caractéristiques : Fusionne les caractéristiques existantes afin de produire de nouvelles, plus représentatives.

Choix des caractéristiques : Isoler et garder uniquement les caractéristiques les plus significatives.

II.4. Représentation des données et clarification des résultats

Pour les opérations de machine learning supervisé ou non supervisé, il est essentiel de posséder des données d'entrée que l'ordinateur peut comprendre. Ces informations sont généralement disposées en format tableau où :

- Chaque ligne illustre une instance (un exemple de données).
- Chaque colonne représente une propriété décrivant cet exemple.

Il est aussi crucial de différencier les diverses sortes de sorties :

- Résultats distincts : Provenant d'un ensemble limité et spécifique de valeurs, elles sont caractéristiques des problèmes de classification.
- Sorties continues : Elles font partie d'un ensemble continu et sont fréquemment trouvées dans les problèmes de régression.

II.5. Synthèse des modèles d'application

Une table récapitulative peut être utilisée pour présenter les contextes d'application des différents types de données dans les algorithmes d'apprentissage supervision et non supervision.

Table 5.1 *: Type de données vs type d'apprentissage [2].*

Type de donnée	Apprentissage supervisé	Apprentissage non-supervisé
Discrète	Classification	Regroupement
Continue	Régression	Réduction de dimensionnalité

II.6. Apprentissage par renforcement

L'apprentissage par renforcement enseigne à un agent comment se comporter de manière optimale dans un environnement spécifique pour atteindre un but fixé par l'utilisateur. On décompose ce genre de problème en une série d'étapes consécutives. À chaque phase, l'agent doit sélectionner une action parmi un ensemble proposé, ce qui facilite son interaction avec l'environnement.

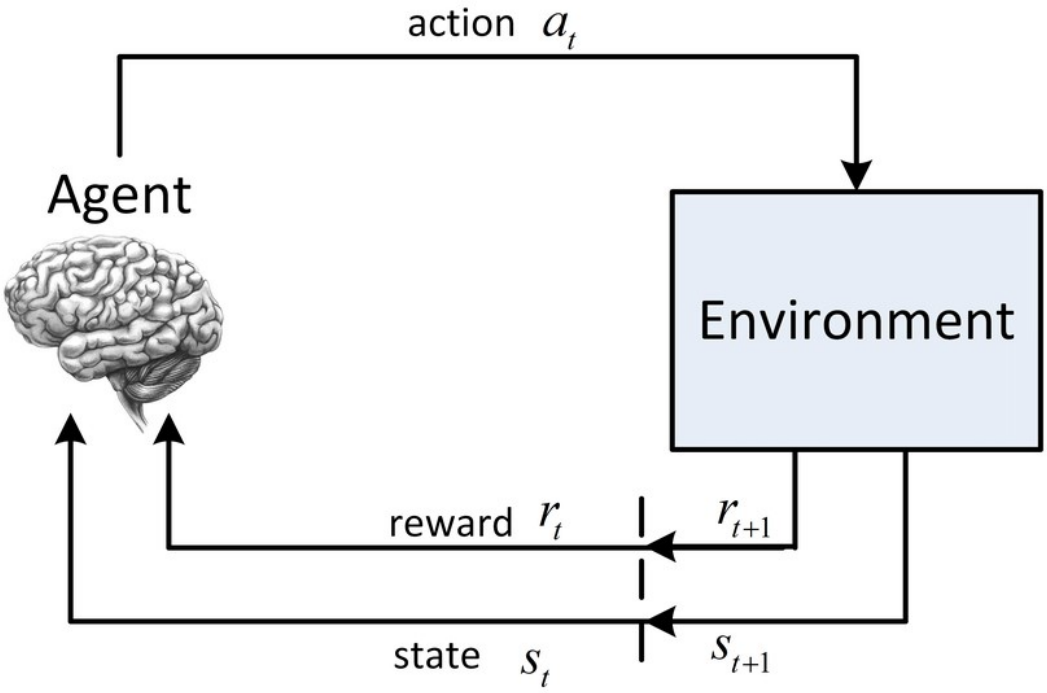


Figure 2.5 : *Interaction agent-environnement*

À l'inverse de l'apprentissage supervisé, aucune directive explicite de comportement cible n'est donnée pour orienter le processus d'apprentissage. Au lieu de cela, l'agent reçoit un signal paramétré par l'utilisateur qui indique si l'action sélectionnée était adéquate. L'agent se sert de ces retours pour ajuster et perfectionner ses prises de décision.

Cette forme de formation est particulièrement appropriée pour une multitude d'applications en robotique. Il se démarque de l'apprentissage supervisé et non-supervisé par l'emploi d'un signal de récompense qui indique simplement la qualité de l'action entreprise par l'agent, sans préciser la meilleure voie à suivre. En outre, il ne se sert ni des données d'apprentissage ni des étiquettes [3].

II.7. NEURONE BIOLOGIQUE

Le cerveau humain est composé d'environ 10^{11} neurones, ce qui représente un trillion, chaque neurone établissant entre 1 000 et 10 000 connexions connues sous le nom de synapses. Un neurone est une cellule dotée d'un corps cellulaire qui joue le rôle de centre de contrôle, où les données reçues sont additionnées (consulter la Figure II.6).

Les dendrites, qui se développent à partir du corps cellulaire, acheminent les données de l'environnement vers le neurone. Suite au traitement, ces données sont acheminées vers d'autres neurones par le biais de l'axone. Le lien entre deux neurones, connu sous le nom de synapse, est crucial pour la diffusion et l'analyse des signaux [3].

Les réseaux neuronaux biologiques ont la capacité d'effectuer des tâches complexes comme la mémorisation, l'apprentissage par exemplification, la généralisation, l'identification de formes et le traitement des signaux. Sur la base de la structure et du fonctionnement des neurones biologiques, les travaux de recherche ont abouti à l'élaboration de neurones formels, souvent dénommés neurones artificiels [3].

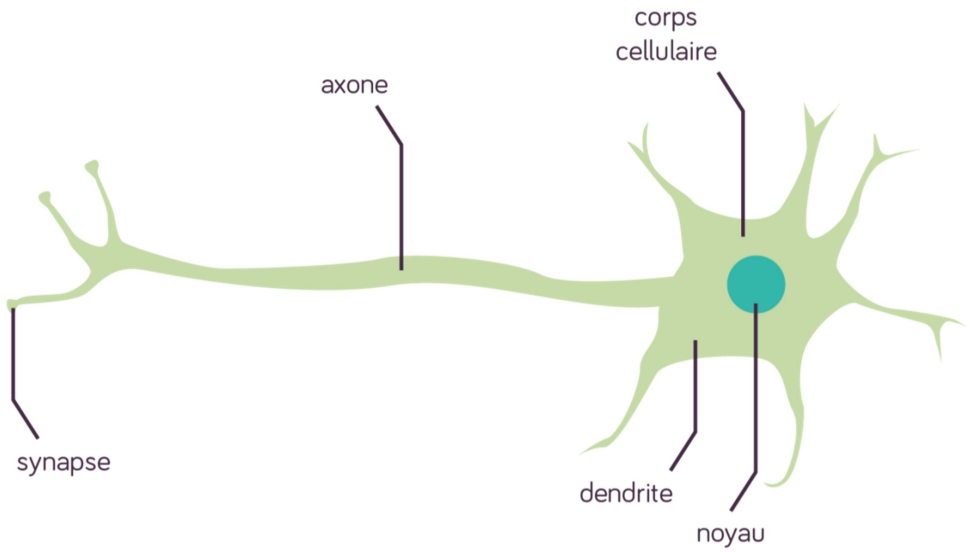


Figure 2.6 : *Le neurone biologique [3].*

II.7.1. LE PERCEPTRON

Le perceptron est le premier réseau de neurones artificiels évolutif, conçu pour apprendre. Initialement, il avait pour objectif de reconnaître les lettres de l'alphabet à l'aide de capteurs photoélectriques.

Fondamentalement, le perceptron repose sur une fonction mathématique. Les données d'entrée (x) sont pondérées par des coefficients (w), et leur produit génère une valeur numérique. Cette valeur peut être soit positive, soit négative. Lorsque cette valeur est positive, c'est-à-dire lorsque le poids total des données d'entrée dépasse un seuil prédéfini, le neurone artificiel s'active [1].

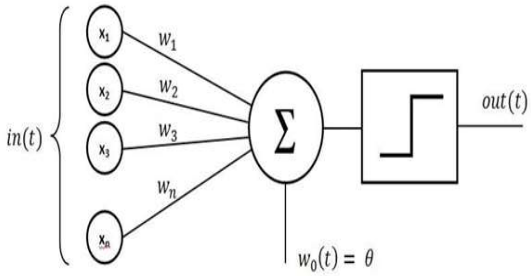


Figure 2.7 : *Réseau monocouche [1]*

II.7.2. PERCEPTRON MULTICOUCHE

Le Perceptron Multicouche, ou Multi layer Perceptron (MLP) en anglais, est le premier réseau de neurones artificiel à avoir été largement appliqué dans diverses pratiques, comme l'identification des fleurs, la détection des fraudes et bien d'autres domaines.

Ce type de structure comporte plusieurs niveaux de neurones, y compris une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque neurone d'une couche a une connexion complète avec tous les neurones des couches adjacentes (précédente et suivante). Chaque liaison est liée à un poids qui définit l'effet de l'activation d'un neurone sur celui des neurones suivants [8].

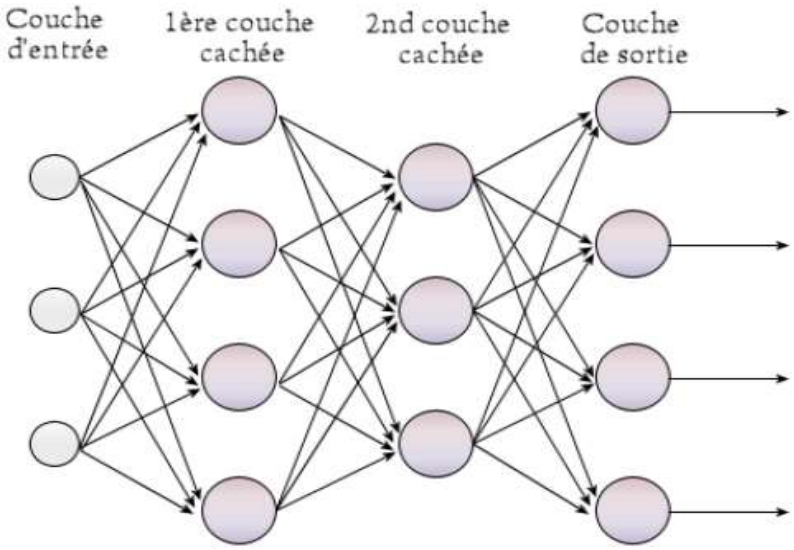


Figure 2.8 : *Perceptron multicouche*

II.8. Coefficient Cepstraux en Fréquence Mel (MFCC)

Les MFCC (mel frequency cepstral coefficients) sont des caracteristiques acoustique tres utilisees dans le domaine du traitment automatique de la parole notamment pour des taches telless que la reconnaissance vocale,l'identification du locuteur ou encore la classification d'emotions. Ils permrttant de representer le contenu spectral d'un signal audio de manière à imiter la perception auditive humaine en particulier la sensibilite de l'oteille aux defferants frequences.

La transformation MFCC en base sur l'idee que l'oreille humaine ne percoit pas les fre quemces de manière lineaire, elle est plus sensible aux basse frequences qu'aux hautes. Pour tenir compte de cela le spectre du signal est d'abord converti en echelle de mel une echelle de frequence non lineaire qui reflete cette perception.

Le calcule des MFCC passe par plusieurs étapes :

- Découpage **du signal en fenêtres (frames)** pour analyser de courts segments du signal audio.
- **Application de la transformée de Fourier (FFT)** sur chaque segment pour obtenir le spectre de fréquence.
- **Filtrage sur l'échelle de Mel**, grâce à une banque de filtres triangulaires répartis selon cette échelle.
- **Calcul du logarithme** de l'énergie de chaque filtre Mel (pour simuler la perception logarithmique humaine).
- **Application de la transformée en cosinus discrète (DCT)** pour obtenir les coefficients cepstraux finaux, qui forment les MFCC.

Ces coefficients permettent de **réduire la dimension du signal tout en conservant l'information pertinente** pour la reconnaissance. Ils sont compacts, robustes au bruit, et très efficaces pour capturer les caractéristiques vocales propres à un locuteur ou un phonème.

II.9. RESEAUX DE NEURONES ARTIFICIELS (RNA)

Un réseau de neurones artificiels (RNA) est un dispositif informatique qui s'inspire du mode de fonctionnement du cerveau humain, et qui est prévu pour être exploité sur des ordinateurs équipés d'intelligence artificielle. Il se base sur l'architecture des neurones biologiques présents dans le cerveau humain.

Les RNA comprennent au moins deux strates de neurones : une couche d'entrée et une couche de sortie. Ils comportent typiquement des strates intermédiaires, également connues sous le nom de couches cachées (hidden layers). Le nombre de couches requis dans le réseau est défini par la complexité du problème à traiter.

Chaque stratum est composé d'un vaste ensemble de neurones synthétiques, chacun dédié à une mission spécifique. Les réseaux de neurones artificiels sont utilisés dans divers domaines tels que la classification, la régression et l'évaluation des densités de probabilité. Ils trouvent également leur application dans l'apprentissage supervisé et non supervisé, de même que dans les modèles discriminants et générateurs.

Pour faire court, les réseaux de neurones artificiels constituent des instruments d'une grande souplesse et efficacité, présentant un potentiel notable pour traiter des problèmes complexes et diversifiés [1].

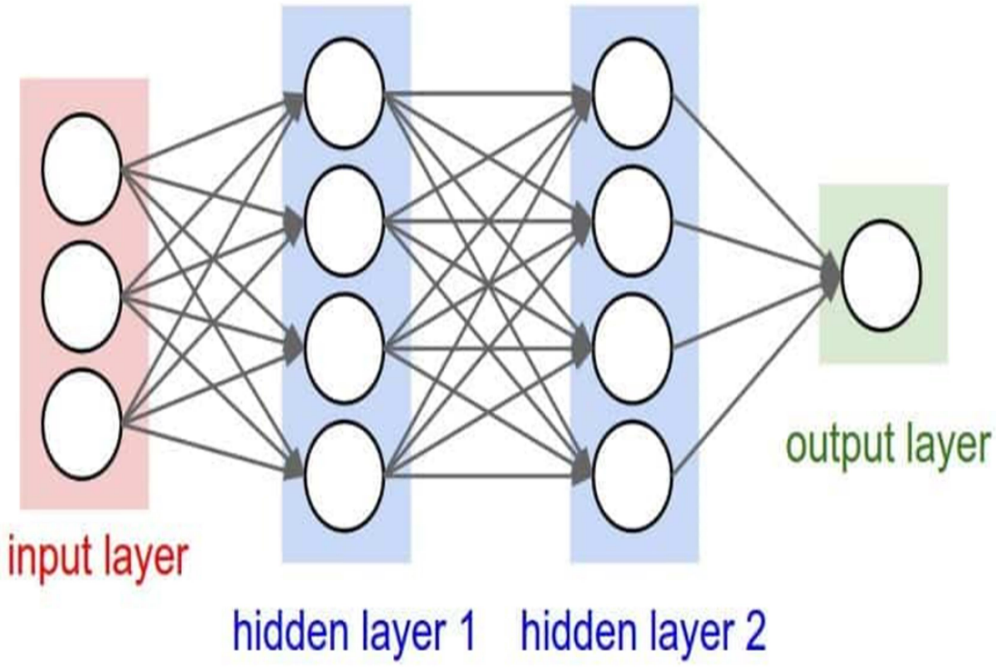


Figure 2.9 *: Réseau de neurones artificiels*

II.9.1. Le Fonctionnement des Réseaux de Neurones Artificiels

Les réseaux de neurones artificiels opèrent par le biais de multiples processeurs travaillant simultanément, disposés en strates.

Couches du réseau :

- La couche initiale, désignée sous le nom de couche d'entrée, traite les données non traitées.
- Chaque niveau ultérieur gère les données provenant du niveau antérieur.
- La couche finale, connue sous le nom de couche de sortie, produit les résultats du système.

Afin de traiter des problèmes complexes, le réseau peut inclure plusieurs niveaux intermédiaires. Chaque neurone a une valeur qui définit les données qu'il est capable de transmettre.

Fonctionnement des neurones

Chaque neurone utilise une fonction d'activation pour déterminer sa valeur de sortie. Cette évaluation sert à établir combien de neurones doivent être stimulés pour résoudre un problème spécifique.

Apprentissage

Un algorithme est élaboré pour lier chaque entrée à une sortie déterminée. Cette méthode offre au réseau la possibilité d'apprendre à partir de nouvelles informations. Le réseau développe la compétence d'exécuter des tâches spécifiques en étudiant des exemples généralement marqués.

Ainsi, les réseaux de neurones artificiels ont la capacité d'apprendre à identifier des objets dans des images, parfois avec une exactitude qui dépasse celle du cerveau humain. Toutefois, à l'instar du cerveau humain, ces réseaux ne se laissent pas programmer de manière directe. Ils doivent se former en examinant et en analysant des exemples [13].

II.9.2 Les types de Réseaux de Neurones Artificiels

On classe généralement les types de réseaux de neurones selon le nombre de couches requises entre l'entrée des données et la sortie finale. Par ailleurs, le choix du type de réseau dépend du

nombre de nœuds cachés contenus dans chaque modèle. On prend aussi en considération le nombre d'entrées et de sorties à chaque nœud [14]. On distingue :

- Réseaux de Neurones à Propagation Avant.
- Réseaux de Neurones Convolutifs
- Réseaux de Neurones Récurrent

II.9.2.1. Réseaux de Neurones à Propagation Avant

Les réseaux neuronaux à diffusion avant, aussi connus sous le terme de Feedforward Neural Networks, constituent la version fondamentale des réseaux neuronaux artificiels. Dans ce genre de réseau, les informations transitent directement des points d'entrée vers les nœuds de traitement, puis vers les points de sortie, sans recours à la rétroaction ni cycles.

Ce modèle (figure II.10) est facile à utiliser et convient aux missions où les relations temporelles ou séquentielles ne sont pas cruciales [12].

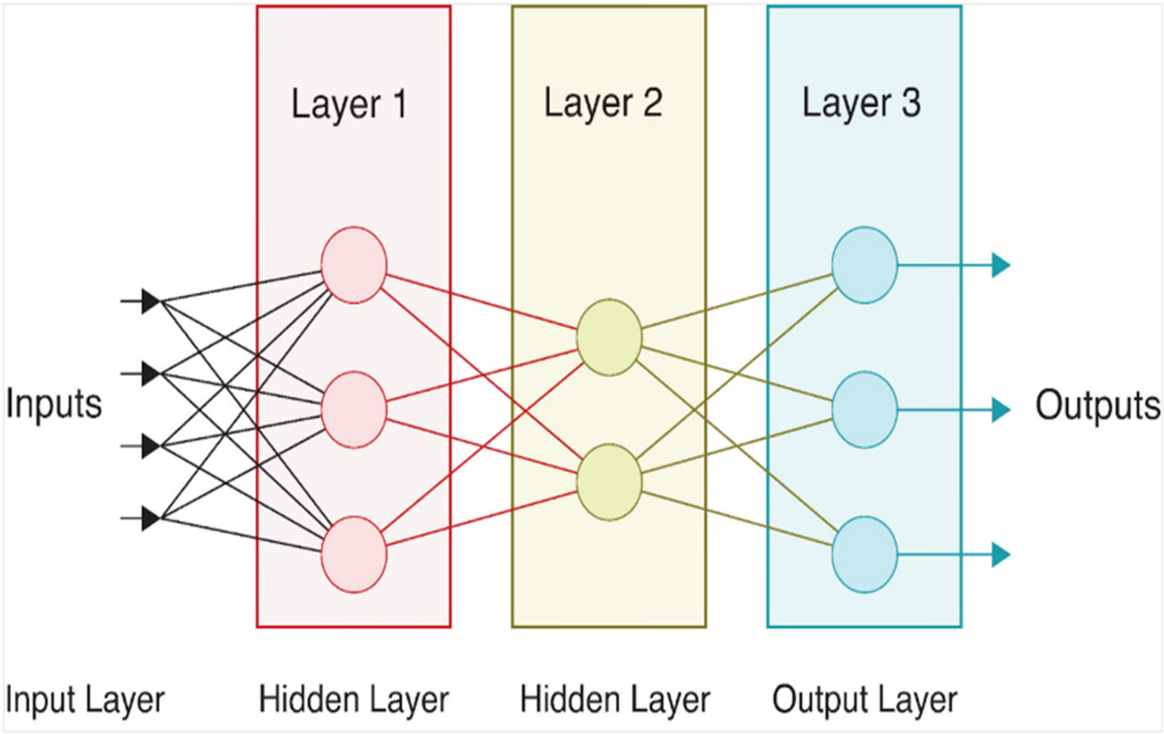


Figure 2.10 : *Réseaux de neurones à propagation avant*

II.9.2.2. Réseaux de Neurones Convolutifs

les réseaux de neurones convolutifs (ou CNN, pour Convolutional Neural Networks) sont spécifiquement élaborés pour examiner des données structurées sous forme de grilles, telles que les images. Ils identifient des patterns simples dans une image en utilisant des filtres de convolution pour déterminer ce qu'elle représente à travers des superpositions successives.

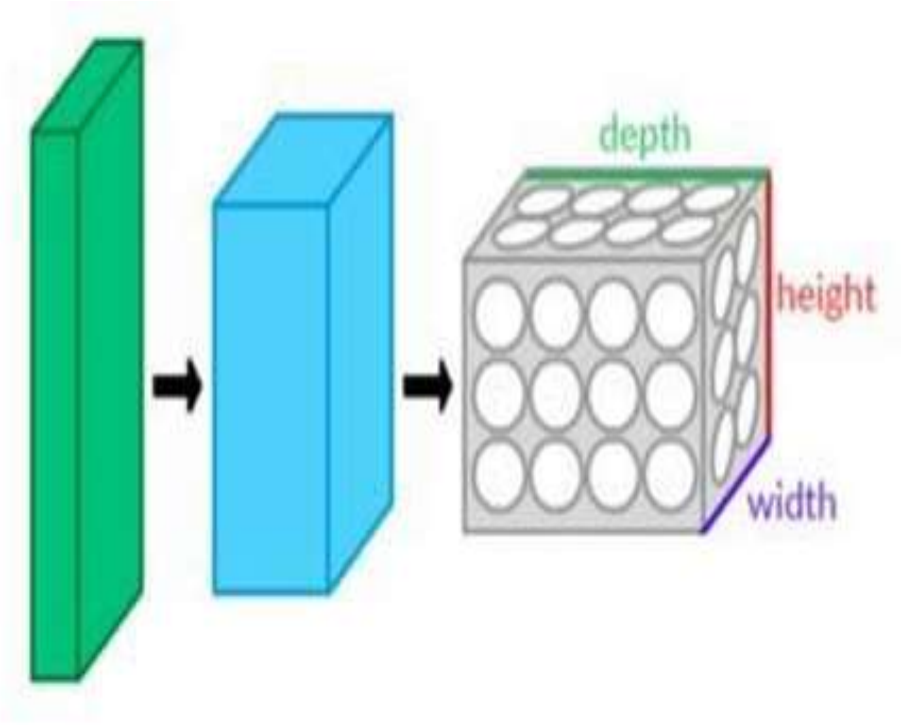


Figure 2.11 : *Réseaux de neurones consolatifs [12]*

Les réseaux neuronaux convolutifs (CNN) sont couramment déployés dans divers secteurs, comme l'identification des visages et la digitalisation de textes. Ils comprennent généralement au minimum cinq strates, avec les résultats issus de chaque strate qui sont transmis aux strates ultérieures, facilitant ainsi une identification graduelle des caractéristiques [13].

II.9.2.3. Réseaux de Neurones Récurrent

Les réseaux de neurones récurrents (Recurrent Neural Networks ou RNN), en français, représentent un type crucial de réseaux neuronaux, généralement employés pour gérer des données séquentielles telles que le langage naturel.

La particularité des RNN est leur aptitude à gérer les séquences tout en considérant le contexte des calculs antérieurs. Dans une séquence, chaque élément est examiné et le résultat dépend des états cachés issus des étapes antérieures, conférant ainsi aux RNN.

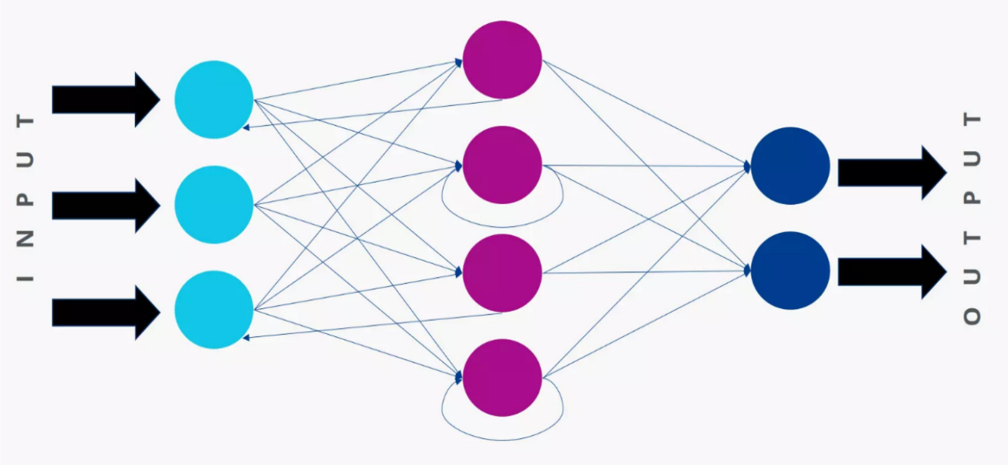


Figure 2.12 : Architecture des réseaux des neurones récurrents

Malgré la capacité théorique des RNN à traiter des séquences de longueur infinie, leur portée pratique est généralement restreinte à des séquences de courte durée, en raison des obstacles associés à la disparition ou à l'explosion du gradient.

Les RNN offrent la possibilité d'exploiter les prévisions antérieures comme entrées pour les phases subséquentes, ce qui en fait un instrument efficace pour modéliser les dépendances dans une séquence [12].

II.9.2.3.1. Les fonctions d'activation dans les RNN

Les fonction d'activation jouent un rôle fondamental dans les réseaux de neurones récurrents (RNN), en introduisant la non-linéarité nécessaire à l'apprentissage de relations complexes entre les éléments d'une séquence. Dans un RNN classique, l'activation la plus couramment utilisée pour pour l'ratât cache est la fonction tangente hyperbolique (tanh). Cette fonction permet de transformer les valeurs linéaires issues des pondération en une plage bornée entre -1 et 1, facilitant ainsi la stabilité de l'apprentissage en limitant l'amplification des valeurs lors de la propagation des états dans le temps mathématiquement, la fonction **tanh** est définie comme suit :

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (1)$$

Elle est particulièrement utile pour centrer les données autour de zéro ce qui améliore la dynamique de l'optimisation lors de l'entrainement du réseau.

En complément de la fonction **tanh** la fonction sigmoïde (ou logistique) est également largement utilisée en particulier dans les architectures de RNN plus avancées telles les LSTM et GRU. La fonction sigmoïde est définie par :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

La fonction sigmoïde avec ses valeurs comprises entre 0 et 1 est idéale pour modéliser les portes dans les unités LSTM et GRU permettant de controles le flux d'information ces potes déterminent quelles données doivent être conservées, oubliées ou transmises facilitant ainsi la gestion des dépendances a long terme. De son cote la fonction tanh est principalement utilisée pour mettre à jour les états caches dans les RNN simples le bon choix ces fonctions d'activation est essentiel pour optimiser l'apprentissage.

II.9.2.3.2. Les type de réseaux de neurones récurrent

Il existe plusieurs variants des réseaux de neurones récurrent (RNN) qui ont été conçues pour optimiser les performances des RNN classiques.

On utilise trios grandes catégories de réseaux neurones récurrents pour gérer de genre de données : Le RNN standard, le LSTM et le GRU.

- Le RNN simple représente la version la plus élémentaire du RNN n'ayant pas de mécanismes pour réguler le passage des informations. Cependant, dans la réalité, l'utilisation des RNN simples est souvent évitée en raison de leurs contraintes.
- Le LSTM représentent une structure de RNN prisée, mise en place pour remédier a la question de l'extinction du gradient. Les LSTM ou réseaux de neurones a mémoire a long terme, représentent une évolution sophistiquée des RNN classiques en intégrantes mécanismes de mémoire a court et long terme . les LSTM font appel a des pour réguler le passage des information et sont performants de dépendances sur le long terme [36]
- Les GRU représentent une autre version du RNN qui end l'architecture des LSTM plus simple en fusionnant les portes d'entrée et d'oubli en une unique porte de mise a jour. Les unités récurrentes generaires (GRU) sont performantes pour des missions comparables aux LSTM, tout en possédant une structure plus simplifiée.

II.9.2.3.3. Architecture de RNN:

Un réseau neuronal récurrent est généralement constitue de plusieurs couches récurrents superposées. Comme indique précédemment, chaque couche récurrente peut être un RNN ordinaire un LSTM ou un GRU.

Un RNN est structuré de la manière suivante :

Entree (input) : le réseau recoit les données séquentielles via son point d'entrée chaque série de données est symbolisée par une suit d'éléments réseau (comme des mots dans une phrase ou des point deans une serie temporelle).

Couches recurrente (recurrente layer) : la couche récurrente est chargée de traite les données séquentielles. Cette couche a des liaisons de rétroaction qui facilitent la transmission des informations d'une phase a l'autre dans la séquence. Ceci offre au réseau la capacite de saisir les dépendances chronologique et d'élaborer des relations complexes entre les éléments de séquence.

Optionnel : on peut superposer plusieurs couches recurrentes constituer un reseau de neurones recurrent profond.chaque couche recurrente gere les donnees issues de la couche anterieure,ce qui permet au reseau de saisir des niveaux de reprentation plus sophistiques et abstraits.

Couche de sortie (output layer) : cette couche produit les predictions ou resultats desires bases sur les donnees traitees par les couches recurrents. Le type de couche de sortie utilise sepend de la nature de tache a accomplir par exemple, dans le cas de la calassifiction , on peut recourir a une couche dense de sortie avec une foncyion adequqte.

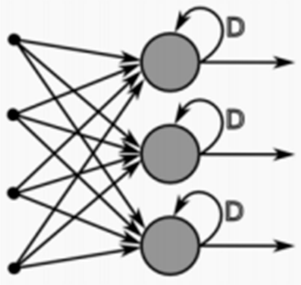


Figure 2.13 : recurrent neural network (RNN) tutorial

II.9.2.3.4. Long Short Terme Mémoire (LSTM)

La mémoire a long court terme (LSTM) est un type particulier de réseau de neurones récurrents (RNN) conçu pour mieux les dépendances a long terme dans les séquences de données. Contrairement aux RNN classique, les LSTM sont capables de retenir l'information sur de longues périodes grâce a une architecture interne composée de cellules de mémoire et portes (entrée, oubli et sortie). Ces portes régulent le flux d'information, permettent au réseau de conserver ou d'oublier certaines données selon leur importance. [33] [35]

Cette capacite rend les LSTM particulièrement efficaces pour des taches comme la reconnaissance vocale, la traduction automatique ou l'analyse de séries temporelles. En filtrant les RNN standards.

Ainsi, les LSTM jouent un rôle crucial dans le traitement du langage naturel et d'autres applications séquentielles.[34]

LSTM utilise trios portes principales, ces portes permettent a l'unite de retenir, d'oublier ou d'ajouter des information a son etat interne, ce qui la rend capable de modeliser des dependances a long terme. Ci-dessous les trois portes principales de l'architecture LSTM :

Porte d'entree (input gate) : elle contrôle quelles nouvelles informations doivent etre ajoutees a l'etat de cellule. Elle est composes de deux parties : une simoide qui selectionne les valeurs a mettre a jour, et une tanh qui cree les nouvelles valeurs candedates.

Porte d'oubli (forget gate) : ce facteur determine quelles donnees de l'etat anterieur doivent etre mises de cote. Elle emploi un fonction sigmoide pour produire un chiffre entere 0 et 1, ou 0 représente effacer totalement et 1 garder entièrement.

Enfin, la porte de sortie (outpute gate) : cela determine la sortie de l'unite LSTM a partir de letat interne actuel. Une fonctin sigmoide selectionne les information a faire sortir, qui sont ensuite modulees par une tanh de l'etat de la cellule.

II.9.2.3.5. Gated Recurrent Unit (GRU)

L'unite recurrent fermeou ou Gated Recurrent Unite (GRU), est variant sipmlifiee du LSTM concu pour traiter le sequences tout en etant plus legere et plus raide a entainer. Elle fusionne les mecanismes de mémoire en combinant la port d'entree et le port d'oubli en une seule port appelee porte de mise a jour. Une auter porte, appeleeporte de reinitialisation,permet de controler l'integration des nouvelles information avec l'etat precedent. 31 32

Cette strecture plus simple reduit le nombre de parametres tout en conservant un capacite efficace a modeliser les dependances a long terme. Le GRU est donc souvent prefere dans des applications en temps reel ou lorsque les ressources de calcule sont limitees. Il donne des performances comparables au LSTM surnombeures taches de traitement du langagenaturel. Grasce asa simplicite, il facilite egalement l'entrainement des modeles sur des ensombles de donnees plus prtits.

Je me suis donc orienté vers un réseau de neurones récurrent (RNN) pour la reconnaissance du locuteur car ce type de modèle est particulièrement adapté pour traiter des données séquentielles,

telles que c'est le cas ici pour la parole. En effet, à la différence de réseaux classiques, qui traitent chaque point de la donnée indépendamment des autres, un RNN prend en compte les relations d'ordre qui existent dans le temps ainsi que le contexte au sein duquel il est utilisé, c'est nécessaire pour modéliser des signaux vocaux.

Conclusion

Dans ce chapitre nous avons explore le domaine d l'apprentissage profond en commençant par une introduction générale a l'intelligence artificielle et a l'apprentissage automatique, tous en soulignant l'importance de l'apprentissage profond et des réseaux de neurones récurrent (RNN). J'ai aborde les bases des réseaux de neurones artificiels en mettant l'accent sur les différentes couches constitutives des RNN et leur role dans le traitement des données séquentielles.

J'ai également mis en lumiere les aventages specifiques des RNN dans le domaine de la reconnaissance du locuteur notamment leur capacite a modeliser les dépendances temporelles et a extraitre des caracteristique partinemtes a partir de donnees audios sequentielles. De plus nous avons discute des architectures et des methode d'entrenement qui contribyent a l'effecacite et a la precision de ces modeles dans ce domaine.

Enfin, j'ai conclu par un analyse comparative enter la prentissage automatique et l'apprentissage profond en mettant en avant leurs distinctions principales et leurs domaines d'application respectifs notamment la capacite des RNN a gerer des donnees complexes.

Chapitre III : Résultats Travaux d'Expérimentation

CHAPITRE III : RESULTATS DES TRAVAUX D'EXPÉRIMENTATION

Introduction

Sont présentés ci-dessous l'ensemble des outils, ressources et choix techniques mis en œuvre pour la conception et la réalisation d'un système d'identification automatique du locuteur par apprentissage profond dans ce chapitre. Pour répondre aux contraintes spécifiques inhérentes au traitement du signal audio dans le domaine du training de modèles d'apprentissage séquentiels, nous avons veillé à sélectionner un environnement de développement approprié et à conserver une structuration rigoureuse des différentes phases expérimentales.

La première partie de celui-ci est consacrée à une description de la plateforme utilisée qui est Python et l'environnement en ligne Kaggle pour son accès aisé à des ressources GPU et plus largement à l'ensemble de l'écosystème constituant un panel riche de bibliothèques scientifiques et de frameworks de deep learning. Dans un second temps, nous expliquons la constitution d'une base de données vocale originale composée de corpus multilingues et de corpus monolingues adaptés à différentes fréquences d'échantillonnage, afin de simuler différents contextes de capture audio.

Par la suite, nous développons le processus d'extraction des caractéristiques acoustiques, réalisé à partir des coefficients cepstraux en fréquence Mel (MFCC), de l'architecture du réseau de neurones récurrents (RNN) utilisé, du choix de l'optimiseur et des paramètres d'entraînement, des techniques de régularisation et des techniques d'augmentation de données.

Enfin, nous faisons état des résultats obtenus lors d'évaluations précises mettant à profit diverses métriques (précision, loss, et matrice de confusion). L'influence des différentes configurations (fréquence, contexte phonologique, technique d'amélioration) et de l'intégration du verbal sur la performance du modèle est discutée, à la lumière des conditions favorable à une meilleure identification du locuteur. Une interface graphique a également été développée pour une mise en œuvre accessible et interactive du système.

III.1. Environnement de développement

Dans ce chapiter, nous presentons les outils, les ressourceset les choix techniques utilises pour la realisation du system d'identification automatique du locuteur. L'environnement de développement a été conçu pour repondre aux exigences de traitement audio, d'apprentissage automatique et d'entrainement d'un reseau de neurones récurrents basé sur les caracteristique MFCC.

III.1.1. Plateforme de développement

Langage python :

Python est l'un des langages de programmation les plus couramment utilisés par les professionnels de la donnée, il a été inventé par Guido van Rossum, la première version de python est sortie en 1991. Ses applications ne sont pas limitées à la Data Science mais peuvent également être utilisées pour développer des logiciels, écrire des algorithmes ou encore gérer l'infrastructure web d'un réseau social

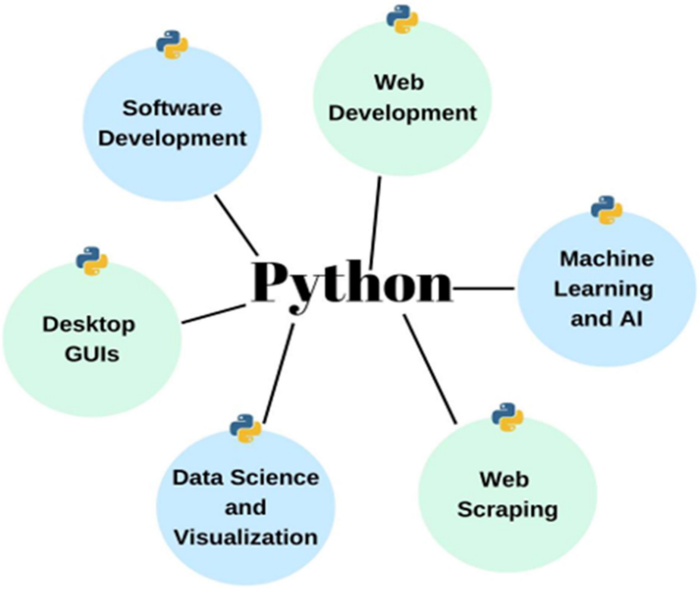


Figure 3.1: les domaines d'application de python

L'ensemble des travaux expérimentaux été réalisé sur la plateforme Kaggle un environnement de développement en ligne permettant d'exécuter du code python dans des notebooks interactifs, avec un accès intègre a des ressources GPU. Ce choix offre un cadre pratique et reproductible pour l'entrainement de modèle de deep learning, notamment dans le domaine de reconnaissance vocal

Kaggle permet également de gérer les bibliothèques nécessaires via un environnement préconfigure, et de stocker les données de mania resécurisée. Son intégration avec trensorflow, scikit-lean, librosa, et d'autres outils d'analyse scientifique, a permis démener l'ensemble des expérimentations de manière fluide

Ce matériel offre une puissance de calcul suffisante pour effectuer des phases de développement locales, notamment le test initial du modèle U-Net sur des images échographiques de taille réduite à moyenne. Cependant, en raison des limites imposées par la capacité de la VRAM (notamment pour des batchs importants ou des résolutions élevées), une partie des entraînements a été réalisée sur la plateforme ***cloud Kaggle,*** qui fournit un accès gratuit à des ressources GPU plus performantes

III.2. Travaux expérimentaux

Dans le cadre de nos recherches expérimentales, nous avons mené une série d'études approfondies qui nous ont permis d'explorer en détail divers aspects de la reconnaissance automatique du locuteur. Le schéma illustré dans la figure décrit de manière détaillée la démarche méthodologique que nous avons adoptée pour notre étude.

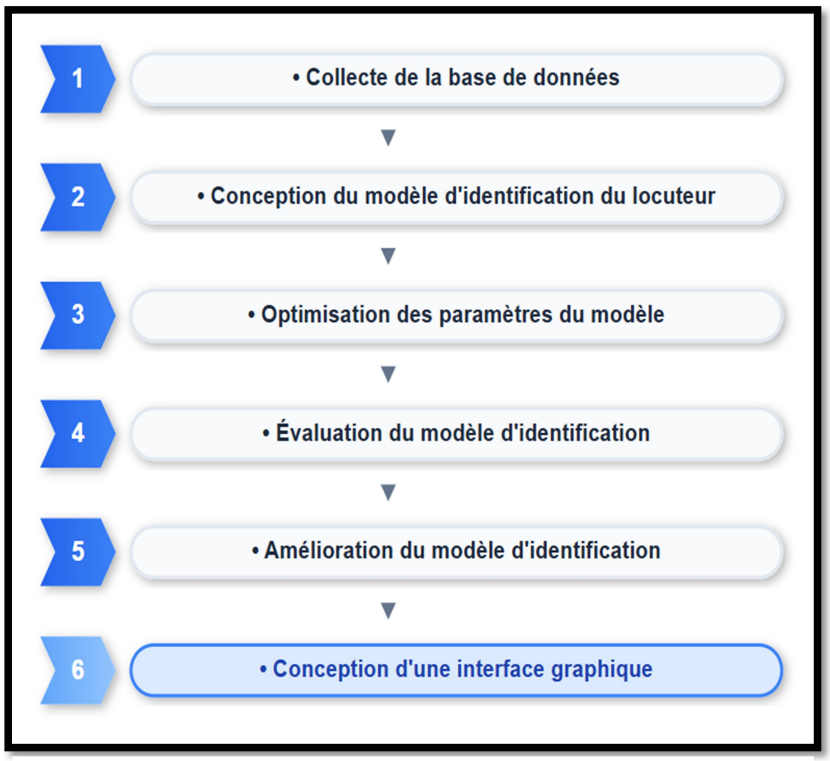


Figure 3.2 : *Démarche méthodologique de notre travail.*

III.2.1. Création de la base de données vocale

La base de données vocale utilisée dans ce projet a été entièrement conçue par l'auteur. Elle comprend les enregistrements vocaux de 25 locuteurs différents, chacun ayant été invité à prononcer un script préétabli. Ce script a été uniformément applique a tous les participants afin de garantir une cohérence linguistique dans le contenu vocal analyse.

Pour chaque locuteur, les enregistrements ont été réalisés a trios fréquences d'échantillonnage déférentes

- 8 khz : fréquences téléphoniques (qualité faible),
- 16 khz : standard utilise pour le traitement de la parole (qualité moyenne),
- 44 khz : qualité audio haute fidélité.

Cette diversité fréquentielle permet d'analyser la robustesse du système face a des variations de qualité du signal et simule des contextes d'enregistrement varies.

Les fichiers ont été organises en sous-dossiers par locuteur et nommes de manière à faciliter leur identification automatique. Tous les enregistrements sont au format .wav, non compresse, assurant une haut fidélité pour l'extraction des caractéristiques acoustiques.

III.2.2. Conception du modèle d'identification du locuteur

Dans cette partie, nous décrivons en détail l'approche méthodologique déployée pour développer notre modèle d'identification, illustrée par l'organigramme représenté dans la **figure III.3**

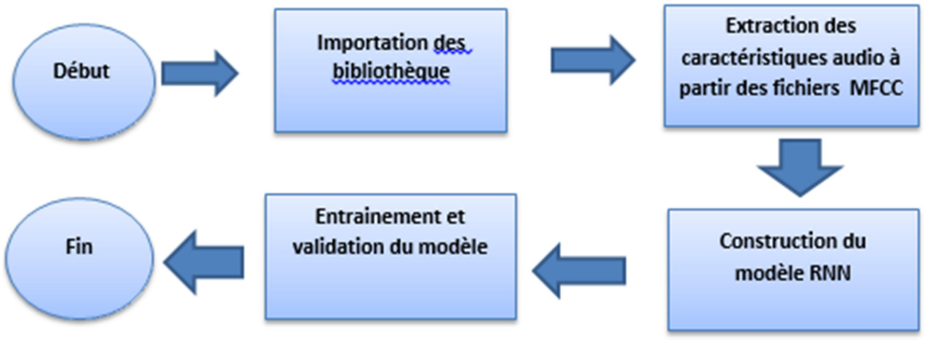


Figure 3.3 : *Modèle d'apprentissage.*

Le processus méthodologique adopté dans cette étude se décompose en quatre étapes principales, qui sont cités dans l'organigramme. Chacune de ces étapes fera l'objet d'une description détaillée dans les sections suivantes.

III.2.2.1. Importation de la bibliothèque

Les différentes bibliothèques utilisées et leurs fonctions sont présentées dans le tableau III.1 :

Table 6.1 : *bibliothèque python utilisée dans le modèle d'identification [2].*

Bibliothèque	**Fonction**
Numpy	Opérations numériques
Tensorflow	Construction et entraînement du modèle de Deep Learning
Librosa	Traitement audio (lecture, analyse)
Os	Gestion des fichiers et des répertoires
Sklearn	Fournit des outils pour l'apprentissage automatique (modèles, évaluation, preprocessing).
Soundfile	Lecture et écriture de fichiers audio (formats WAV, FLAC, etc.).
matplotlib	Visualisation des données sous forme de graphiques et de diagrammes.
streamlit	Création de l'interface web interactive

III.2.2.2. Extraction des caractéristiques audio

Cette étape comprend les opérations suivantes :

a. Chargement des fichiers audio à l'aide de librosa

b. Calcul des MFCC (Mel-Frequency Cepstral Coefficients) qui représentent les caractéristiques spectrales du signal audio

III.2.2.2.1. Chargement et préparation des données

Le Processus de Préparation des Données consiste à :

1. **Parcours des répertoires d'étiquettes**
   - ✓ Exploration systématique de l'arborescence des dossiers
   - ✓ Association automatique des noms de répertoire aux étiquettes de classification
   - ✓ Détection des sous-dossiers contenant les échantillons audios

III.2.2.2.2. Extraction des caractéristiques MFCC

Cette étape comprend les opérations suivantes :

- ✓ Chargement des fichiers audio à l'aide de librosa
- ✓ Calcul des MFCCs (Mel-Frequency Cepstral Coefficients) qui représentent les caractéristiques spectrales du signal audio
- ✓ Filtrer et stocker les caractéristiques extraites dans des listes

III.2.2.3. Construction du modèle RNN

La construction du modèle RN consisté à :

- ✓ Utilisation des couches LSTM pour le traitement des séquences.
- ✓ Application du Dropout pour la régularisation
- ✓ Compiler le modèle avec une fonction de perte : categorical cross-entropy et l'optimiseur : ADAM

III.2.2.4. Entraînement et appréciation du modèle

Dans cette partie nous exposons la démarche de d'entraînement et validation du modèle.

1.Entrainement

Le modèle est entraîné en utilisant les données d'apprentissage on précise plusieurs paramètres d'entraînement parmi lesquels :

- ✓ **Valid split :** cette variable illustre la fraction de données utilisées comme jeu de validation lors de l'apprentissage de notre part, 15% sont des données destinées à la validation et 15% pour l'évaluation.
- ✓ **Sample rate :** la variable ''sample rate'' indique la fréquence d'échantillonnage des enregistrement sonores présents dans notre base de données. Cela représente la fréquence à laquelle le signal audio été capturé.
- ✓ **Batch size :** cette variable détermine la dimension des (batches) employée lors de l'entraînement du modèle j'ai employé un ensemble de : 64.
- ✓ **Epochs :** la variable 'epoches' détermine le nombre d'epochs (cycles complets) au cours desquels le modèle a été entrainé. Une epoche se réfère a un passage intégral sur l'intégralité des données d'apprentissage dans notre code j'ai utilise un total de 100 époques.
- ✓ **Fonction d'activation :** l'outil athématique qu'est la fonction de pette évalue la différence entre les prédictions générées par le modèle et les valeurs cibles réelles. Elle est utilisée comme critère d'optimisation au cours de l'apprentissage orientant la modification des poids du réseau. Un faible taux de perte signifie que le modèle effectue des prédictions précises on identifie deux types : tangente hyperbolique, sigmoïd.
- ✓ **Optimiseur** : l'optimiseur est un algorithme qui ajuste les poids du modèle afin de minimiser la fonction de perte des exemples courants incluent **Adam** et **Nadam** son objectif est de minimiser l'erreur de prédiction sur les données d'entrainement tout rn évitant le surapprentissage.

2.Appréciation du modèle

Une fois l'entraînement terminé, le modèle est évalué sur les données de test afin de mesurer sa performance sur des exemples non traités auparavant. On distingue :

- ➢ **Evaluation :** les indicateurs de performance (comme la précision, l'erreur, la matrix de confusion, etc.) sont déterminés a partir des prédictions du modèle sur le jeu de test.
- ➢ **Présentation des résultats de précision :** un récapitulatif des résultats est fourni soulignant la précision générale du modèle la précision est déterminée comme le correctement classes et du total d'échantillons pressant dans l'ensemble de validation, ce qui en fait un indicateur essentiel pour juger l'efficacité d'un modèle de classification.
- ➢ **Prédiction :** le modèle réalise des prédictions sur les données d'essai. On met en parallèle les classes prévues avec les classes effectives pour démontrer la précision de la classification.

III.2.3. Optimisation des paramètres du modèle

Pour optimiser les performances de notre modèle, nous avons mené plusieurs expérimentations étudiant les phénomènes de sous-apprentissage et de surapprentissage sur l'ensemble des enregistrements en dialecte algérien à 16 kHz. Les résultats présentés dans la figure III.4 illustrent la précision et la perte de l'entrainement et la validation issus par notre script, montrent notamment un sur-apprentissage important avec des taux de :

- ✓ Précision de validation variable, atteignant une valeur maximale de 58,73%
- ✓ Perte de validation qui varie considérablement, avec une valeur minimale qui vaut 2.
- ✓ Test égal à : 64,87% qui est insuffisant pour un système de classification

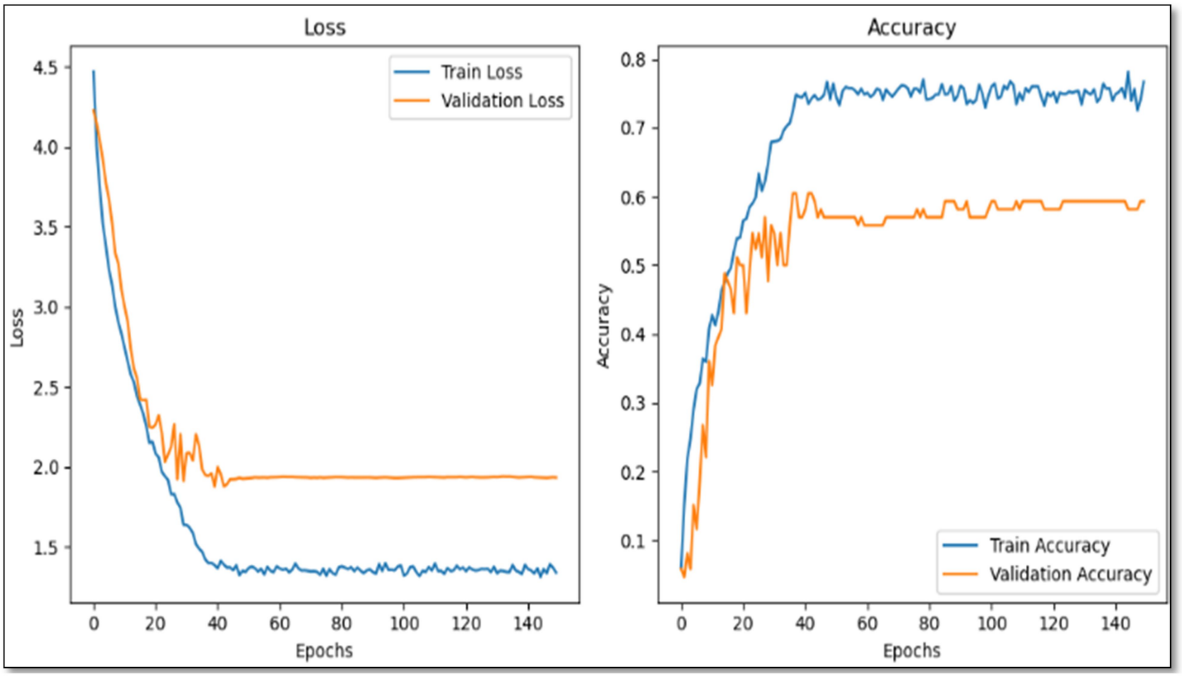


Figure 3.4 : *Evaluation de la précision et perte pour l'entrainement et la validation.*

L'optimisation de notre script consiste à utiliser un algorithme nommé RANDOM SEARCH. Il sert à identifier les paramètres appropriés pour optimiser les taux de précision et de perte en validation, y compris celui de la classification.

Nous avons également intégré d'autres paramètres qui n'étaient pas présents dans le programme initial, à savoir :

Table 3.2 : *les utiliser pour évaluation du modèle*

Technique	Description
Régularisation L1 et L2	Permet d'éviter le surapprentissage en pénalisant les poids excessifs dans le modèle
Shuffle des données	Mélange aléatoirement les données pour empêcher l'apprentissage de motifs artificiels lies à l'ordre.
Taux d'apprentissage adaptatif	Ajuste dynamiquement le taux d'apprentissage pour eviter le sur-apprentissage et le sous- apprentissage
Early stopping	Interrompt l'entrainement automatiquement lorsque la performance sur les données de validation cesse de s'ameliorer.

III.2.3.1. Evaluation du modèle d'identification

Dans cette expérimentation, nous avons appliqué notre modèle à six ensembles d'enregistrements issus de notre base de données, comprenant des données multilingues et monolingues en dialecte algérien, chacune avec trois fréquences différentes : 8 kHz, 16 kHz et 44 kHz. L'évaluation des performances est évaluée par les : taux de perte (Learning et validation) , taux de précision (Learning et validation) et la matrice de confusion avec le taux de test.

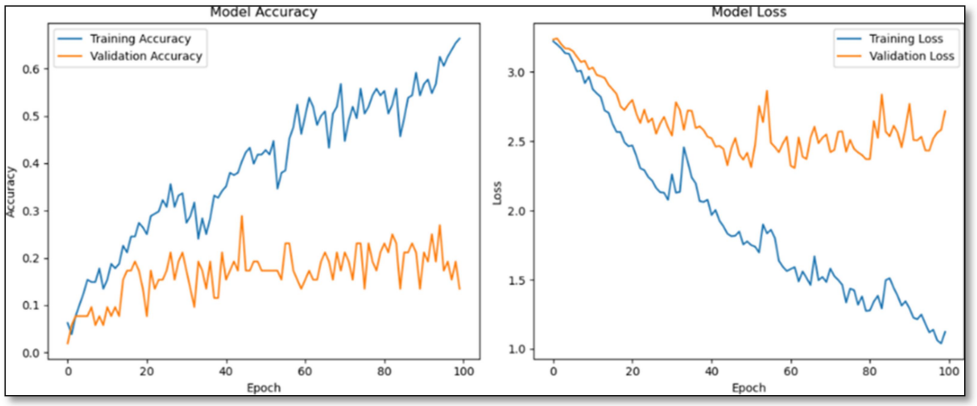


Figure 3.5 : *Evaluation de la précision et perte pour la fréquence 8KHz.*

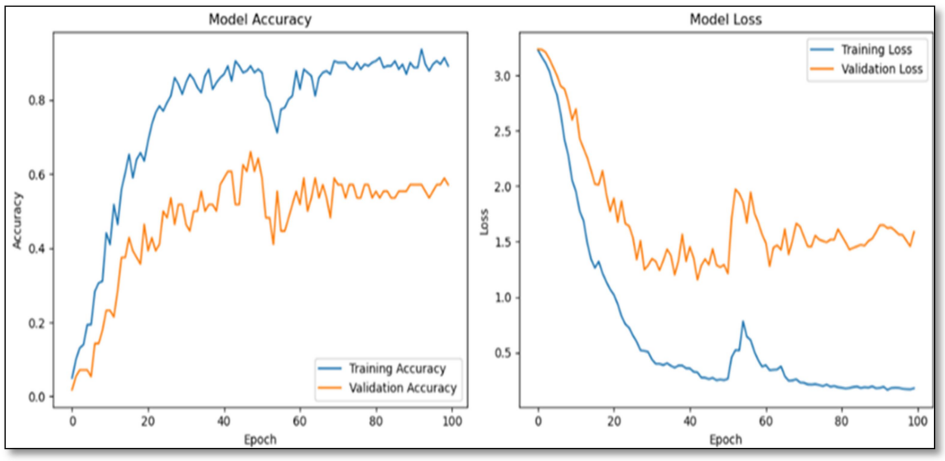


Figure 3.6 : *Evaluation de la précision et perte pour la fréquence 16KHz.*

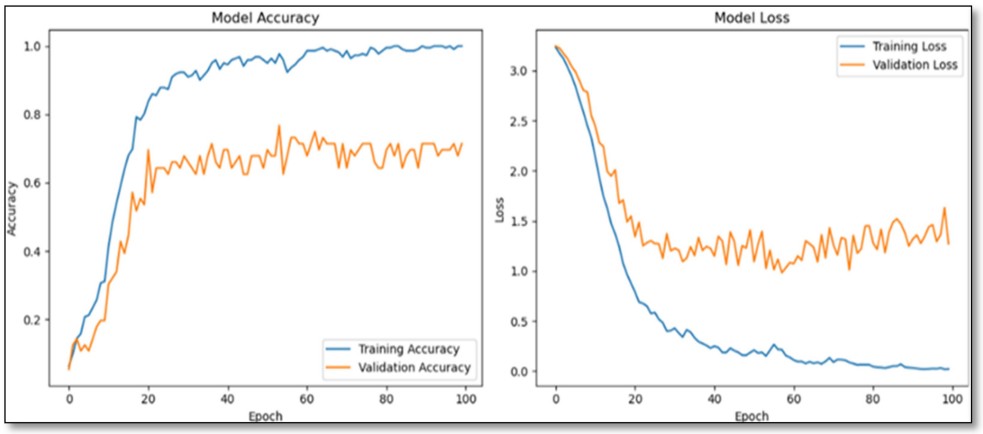


Figure 3.7 : *Evaluation de la précision et perte pour la fréquence 44KHz.*

De manière générale, la précision augmente au fur et à mesure que le nombre d'époques croît, ce qui se traduit par un modèle mieux ajusté au cours de l'entraînement.

Les données récoltées montrent que l'ensemble des enregistrements multilingues présente les performances les plus élevées, ce qui montre qu'il est robuste face aux éléments de traits vocaux. Plus précisément celui à 44 KHz de fréquence d'échantillonnage. Toutefois, tous les taux de précision validée ainsi que le taux de classification restent trop peu élevés pour un modèle d'identification, qui nécessite un minimum de 90% de seuil.

III.2.3.2. La matrices de contusion

La matrice de confusion et le taux de test On appelle matrice de confusion, la matrice de performance qui permet d'analyser les performances d'un modèle de classification. C'est un objet d'évaluation qui permet d'opposer les prévisions de notre modèle aux valeurs réelles dont nous disposons. Cette matrice affiche les différents cas que peuvent prendre une prédiction : le nombre de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs. Nous classons les valeurs dans une matrice à double entrée dont les dimensions sont les classes réelles en ligne et les classes prédites en colonne :

- Chaque ligne correspond ainsi à l'identité réelle (le « vrai locuteur » ou « vraie personne »).
- Chaque colonne est dédiée à l'identité qui a été prédite par ton système.
- Les valeurs sur la diagonale (de haut gauche à bas droite) sont les bonnes prédictions.
- Les valeurs hors diagonales sont les erreurs (confusion entre identités).

La matrice de confusion à 8 kHz monolingue est moins performante que les matrices de confusion à 16kHz et 44 kHz, car ce dernier a tour à tour plus de caractéristiques propres à la reconnaissance du locuteur. Les matrices de confusion multilingues dans les langues présentent également les meilleurs résultats, car l'argument de la langue est mieux pris en compte.

En matière de classification des locuteurs au sein de la matrice de confusion monolingue codée à 8 kHz, des résultats bien nettement moins performants sont observés pour celles codées à 16 kHz et à 44 kHz, tant les enregistrements effectués à ces fréquences possèdent les atouts spécifiques à la reconnaissance du locuteur.

En ce qui concerne les matrices de confusion multilingues, elles aussi affichent, pour les mêmes raisons déjà évoquées, un meilleur score.

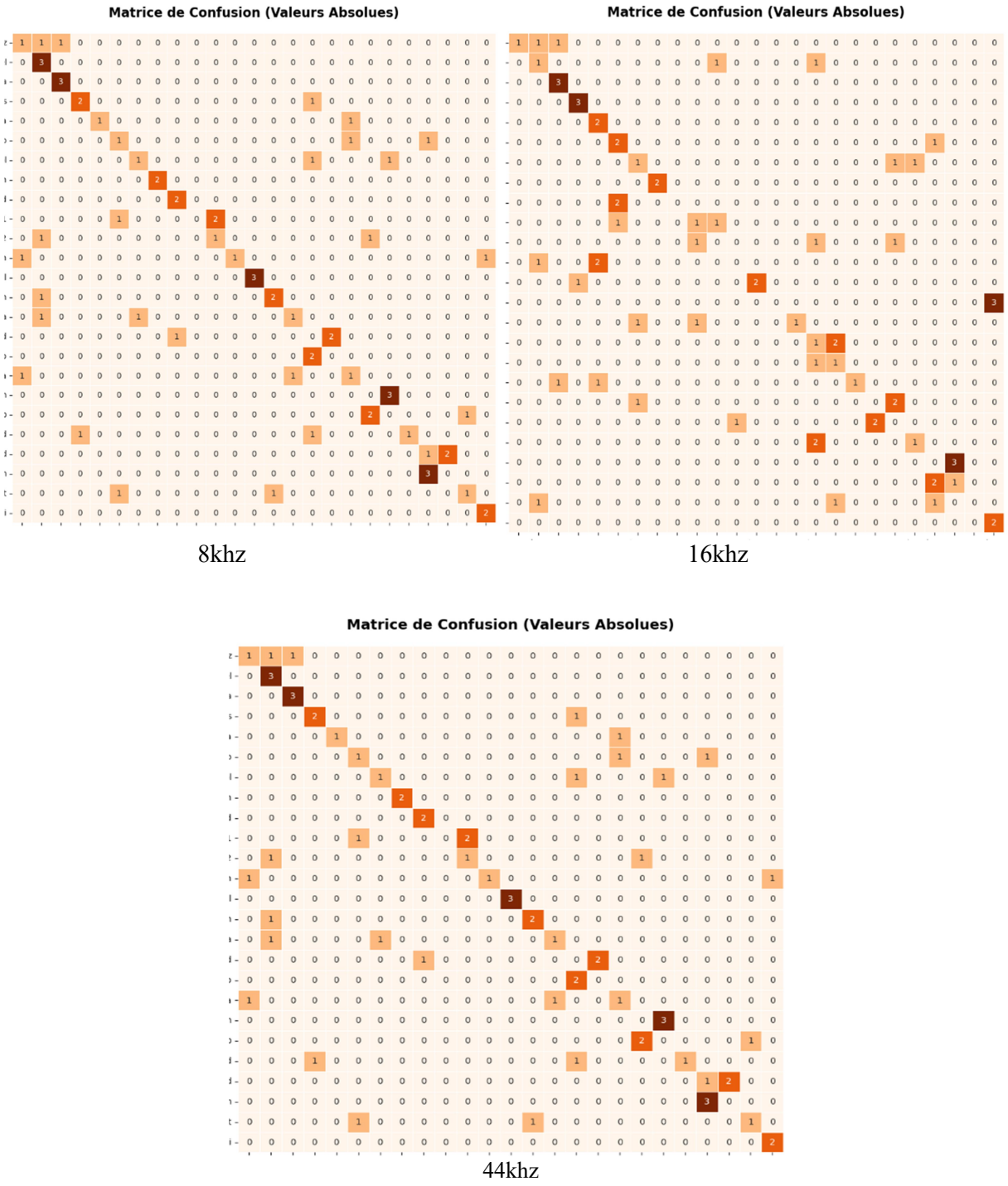


Figure 3.8 : *les Matrice de confusion avant améliorée (8khz, 16 Khz et 44 Khz).*

3.2.4. Amélioration du modèle d'identification

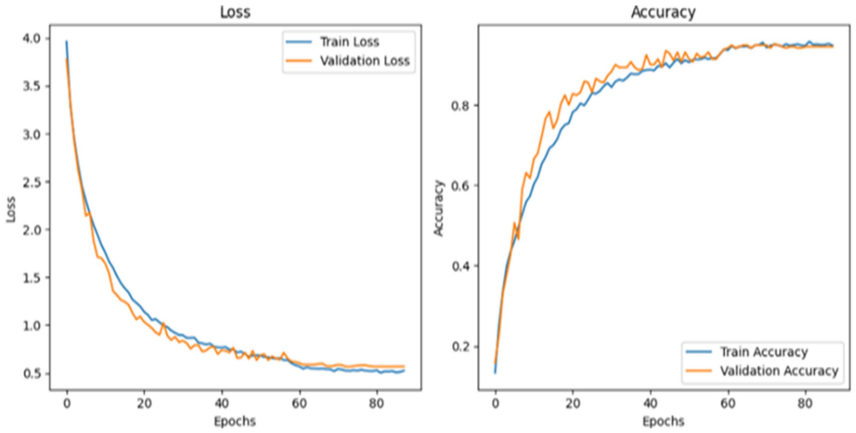
Afin de renforcer la performance de notre modèle d'identification, nous avons eu recours à l'augmentation de données (Data Augmentation), preuve de l'efficacité de cette technique. Effectivement, un des objectifs de l'augmentation de données est de développer le nombre de données tout en cultivant la qualité et la diversité de notre base de données. Pour ce faire, nous avons écrit un script d'application de transformations variées sur les enregistrements actuels, ces transformations sont :

- La vitesse de lecture,
- Le bruit de fond,
- La tonalité,
- Le réglage du volume (mis en œuvre sur deux amplitudes différentes).

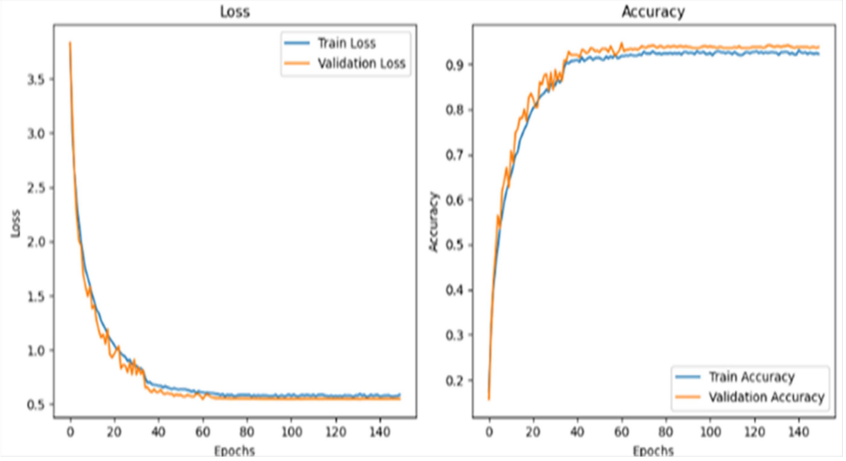
Ces modifications ont permis de multiplier le nombre d'enregistrements par cinq, d'où un enrichissement de la base de données.

À la suite de l'augmentation des données, on constate une nette amélioration des performances du modèle, de la qualité de la précision et de sa perte significativement réduite. On obtiendra alors des matrices de confusion plus performantes entre locuteurs. Ce qui va démontrer l'impact favorable de l'enrichissement de données sur la qualité de classification.

Les résultats révélés dans cette étude démontrent l'effet notable de la fréquence d'échantillonnage et du contexte langagier sur les performances du modèle puisque, en situation monolingue, les taux de précision atteignent 91 % en validation et 94 % en test à 8 kHz, 93 % en validation et 97 % en test à 16 kHz et 94 % en validation et 96 % en test à 44 kHz, alors qu'en contexte multilingue, les performances sont tout à fait similaires mais légèrement supérieures 90,54 % en validation et 94 % en test à 8 kHz, 94 % à 16 kHz autant en validation qu'en test puis 95 % tant en validation qu'en test à 44 kHz. Par ailleurs, l'augmentation de la fréquence d'échantillonnage couplée à un traitement multilingue améliore considérablement la précision du modèle.



8khz



16khz

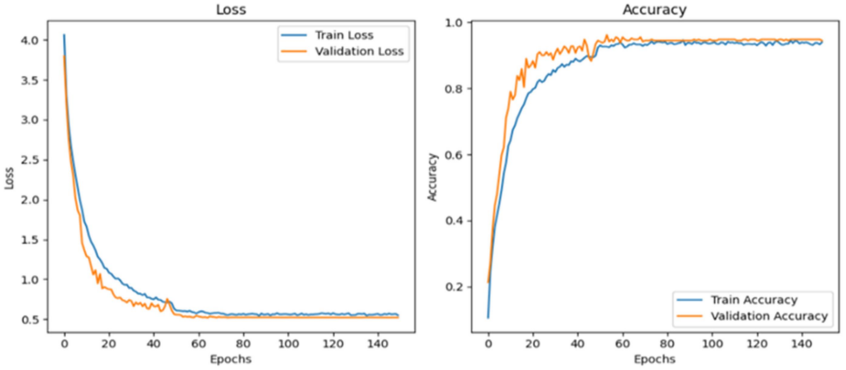
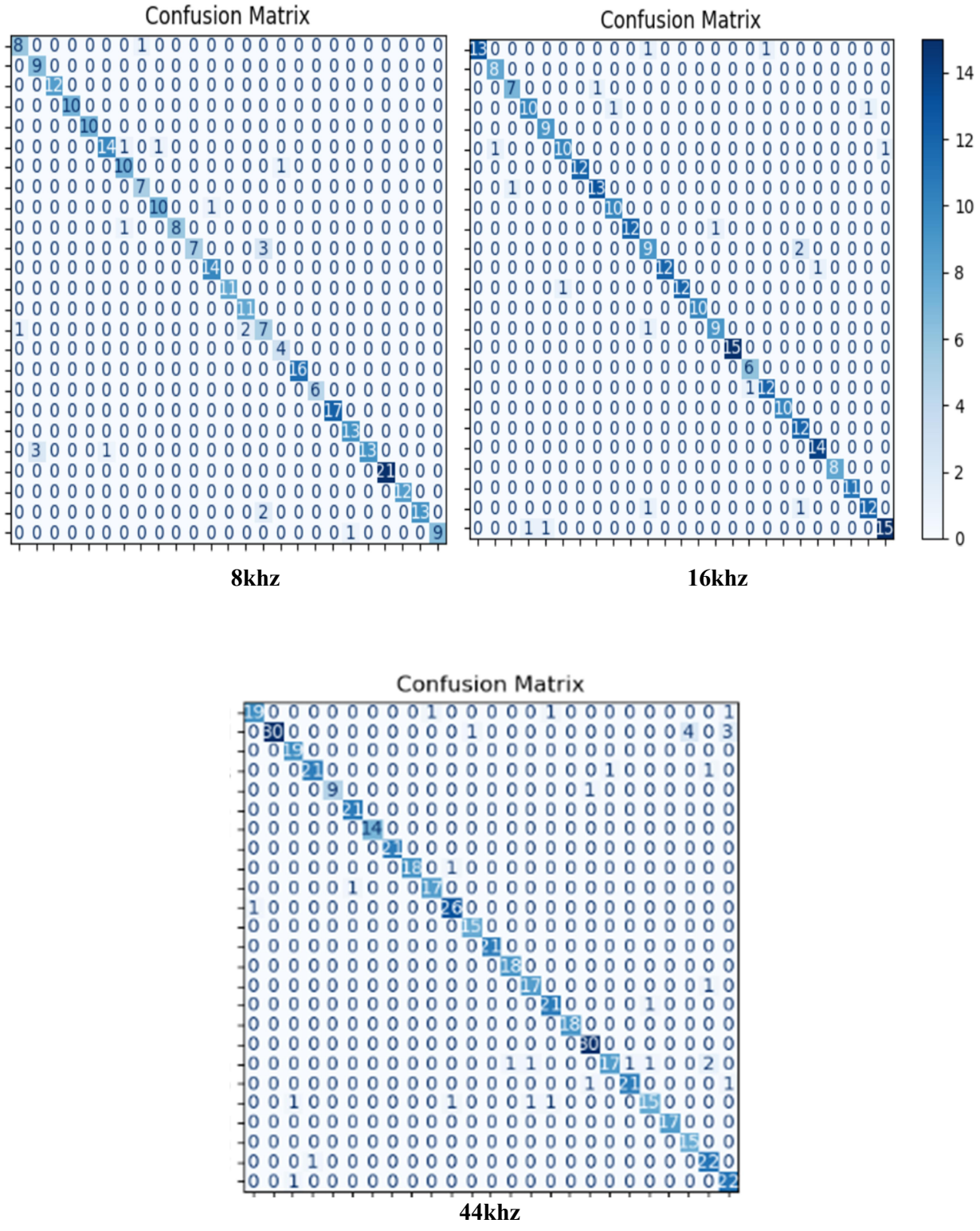


Figure 3.9 : *Pertes et précisions de validation améliorées (8khz, 16 Khz et 44 Khz) .*



8khz **16khz**

44khz

Figure 3.10 : *les Matrices de confusion améliorées (8khz, 16 Khz et 44 Khz) .*

III.2.5. Interface graphique

Afin de faciliter l'accès au système pour les utilisateurs, une interface graphique conviviale a été développée, offrant deux modes d'utilisation (figure III.11) : l'enregistrement vocal en temps réel et le téléchargement d'un enregistrement préexistant. Pour utiliser la première méthode, il suffit de cliquer sur le bouton « Tester en direct », ce qui permet d'activer le microphone et de procéder à l'identification du locuteur. La seconde méthode consiste à sélectionner le bouton correspondant au test via un enregistrement existant. Dans ce cas, l'utilisateur peut importer un fichier audio en saisissant le chemin d'accès ou le lien de l'enregistrement vocal.

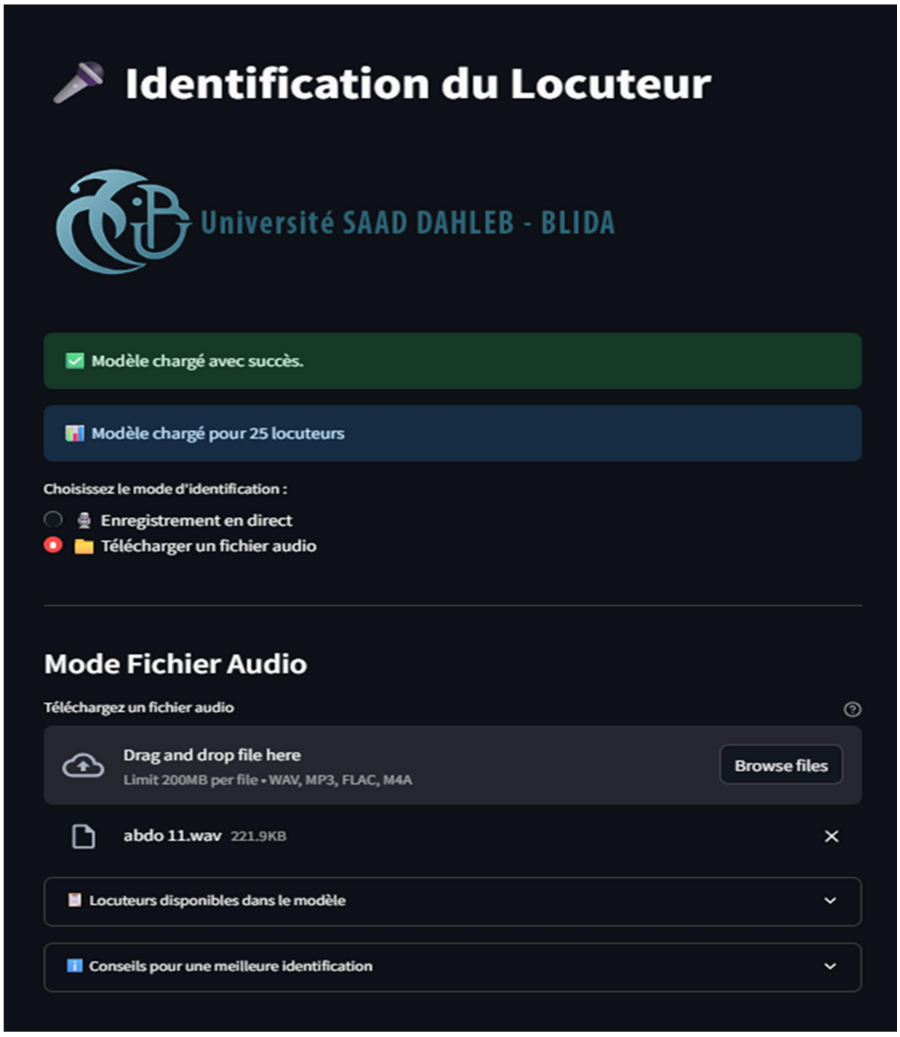


Figure 3.11 : *Interface graphique de notre modèle d'identification.*

Conclusion :

Au sein de ce chapitre, nous avons pu présenter les différentes étapes de sa conception et sa phase d'expérimentation d'un système d'identification automatique du locuteur utilisant le deep learning. L'environnement utilisé a été décrit, à savoir la plateforme Kaggle ainsi que le langage Python, tout comme les choix technologiques mis en œuvre pour élaborer le système. En termes de résultats, le modèle présente un bon niveau de performance avec des taux de précisions autour de 95 % en contexte multilingue et avec une fréquence d'échantillonnage élevée. En outre, l'augmentation de données a également permis d'améliorer fortement ses performances. Une interface graphique a également été élaboré pour faciliter son utilisation. Nous avons pu souligner ici l'impact de la fréquence d'échantillonnage et du contexte langagier sur les performances du modèle au travers de cette étude. Les perspectives de ce travail sont l'amélioration continue du modèle et son application en contexte réel. Les résultats sont d'ores et déjà encourageants et permettent d'envisager de futures recherches. Enfin, le système développé peut être mobilisé en grande partie dans des contextes tels que la sécurité, la biométrie ou la reconnaissance vocale. Quant à l'interface graphique elle-même, son élaboration permet d'ores et déjà son utilisation par des utilisateurs non-initiés. C'est ainsi que l'étude contribue depuis ses débuts à la recherche dans le vaste domaine de la reconnaissance vocale et de l'identification du locuteur. Les résultats montrent que le système mis au point peut avoir une réelle application pratique. En effet, les perspectives futures sont de continuer à améliorer le système et d'expérimenter son efficacité en situation réelle. Le système pourrait donc participer à l'amélioration de la sécurité ou à des fins biométriques. L'interface graphique de présentation du système séduira un utilisateur même novice.

CONCLUSION GENERALE ET PERSPECTIVES

Dans ce mémoire, nous avons présenté un système d'identification automatique du locuteur utilisant l'apprentissage profond, en ayant développé un modèle de reconnaissance vocale permettant l'identification du locuteur à partir des caractéristiques vocales de sa voix, à partir d'une base de données vocale originale constituée de corpus multilingues et monolingues de différentes fréquences d'échantillonnage.

Les résultats des expérimentations montrent que ce modèle permet d'atteindre un bon niveau de performance (95%) en contexte multilingue et avec une fréquence d'échantillonnage élevée. L'augmentation de données a également permis d'améliorer fortement les performances du modèle. Une interface graphique conviviale a été développée afin de faciliter l'utilisation du système. Nos travaux ont permis de mettre en évidence l'impact de la fréquence d'échantillonnage et du contexte langagier sur les performances du modèle. Les perspectives de ce travail sont l'amélioration continue du modèle ainsi que son application en contexte réel. Les résultats sont jusqu'ici encourageants et permettront de futures recherches.

Ce travail expérimental nous a permis de mieux comprendre les subtilités de l'identification vocale en contexte multilingue et ouvre de nombreuses perspectives. À l'avenir, il serait intéressant d'explorer l'intégration de modèles plus complexes, comme les réseaux neuronaux convolutionnels (CNN), afin de consolider les résultats obtenus.

En résumé, ce mémoire a contribué à l'avancement de la recherche dans le domaine de la reconnaissance vocale et de l'identification du locuteur. Le système développé peut être utilisé dans des contextes tels que la sécurité, la biométrie ou la reconnaissance vocale. L'interface graphique développée facilite l'utilisation du système par des simples utilisateurs au service de la biométrie vocale.

BIBLIOGRAPHIE

[1] A .Amehray,[Rehaussement de bruitage perceptuel de la parole], thèse de doctorat, école nationale supérieure des télécommunications deBretagne, 2009.
S.K.Singh,P.C.Pandey,«FeatuesandTechniqueForSpeakerRecognition»,Seminar Report, page 5,6,Novembre 03.

[2] M. MOUSS Mohamed Djamel, « Intégration D'un Module De Reconnaissance De La Parole Au Niveau D'un système Audiovisuel – Application Téléviseur », thèse de doctorat, Université Batna 2, AVRIL 2021.

[3] M.DenisJouve,«Reconnaissancedulocuteurenmilieuxdifficiles»,thèsededoctorat,

[4] UNIVERSITÉD'AVIGNONETDESPAYSDEVAUCLUSE,18Juillet2017

[5] Y. AZIZA, « modélisation Ar et arma de la parole pour une vérification robuste du locuteurdansunmilieubruitéenmodedépendantdutexte»,MémoiredeMagister,Université Ferhat Abbas, Sétif, 2013.

[6] H. Satori, M. Harti and N. Chenfour, « Système de Reconnaissance Automatique del'arabe basé sur CMUSphinx », mémoire master, Dhar Mehraz Fès Morocco.

[7] OthmanLachhab,«ReconnaissanceStatistiquedelaParoleContinuepourVoix Laryngée etAlaryngée », Université Mohammed V de Rabat (Maroc), 2017.

[8] Vincent Jousse, « Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription », mémoire master, Université du Maine, 2011.

[9] MichaelFMcTear,«Spokendialoguetechnology:towardtheconversationaluser interface. Springer Science & Business Media », article, page 3, 2004.

[10] M. A. Wissmann et K. M. Béring, « Automatique Language Identification », Speech Communication, article, page 4, 2001.

[11] Mr.Haddab,«reconnaissanceautomatiquedulocuteurparlaméthodedutauxpassage par zéro », mémoire master, université Mouloud mamri de Tizi-Ouzou, 2007/2008Jin,Minho,andYoo,ChangD,«SPEAKERVERIFICATIONANDIDENTIFICATION»,KorêInstitutAvancédesSciencesetTechnologies,RépubliquedeCorée,2004.

[12] SiwarZRIBIBOUJELBENE,«IdentificationduLocuteurparSystèmeHybrideGMMS»,

thèse, TUNISIA, March 22/26/2009.

[13] Dr.ClintSlatton,«ASpeakerVerificationSystem»,thèse,UniversitédeFlorida,2006.

[14] A. Preti, « Surveillance de reseaux professionnels de communication par la reconnaissance du locuteur », Thèse, Université d‟Avignon et des Pays de Vaucluse, France, 2008.

[15] (consulté le 12/06/2023), disponible sur : https://www.editionseni.fr/open/mediabook.aspx?idR=f6e7a7353a3574180124387fa03fdcl,

[16] AmineAbdaoui,«MachineLearning »,article,page5, 1/7/2019.

[17] La Ryax Team, « Deep learning : comprendre les réseaux de neurones artificiels (artificial neural networks) », article, page 3, 2020.

[18] Dr. Ouarda ZEDADRA, « Système de prédiction de la consommation d'énergie basé Deep Learning », Mémoire master, Université de 8 Mai 1945, Septembre 2021.

[19] Pr.BILAMIAzeddine,«ApprentissageIncrémental&MachinesàVecteursSupports

[20] »,UniversitéHADJLAKHDAR–BATNA,18/12/2013

[21] Guillaume Saint-Cirgue,«Apprendre la machine learning en une semaine», 2019.

[22] Houcine Noura & Khelifa Nadia, « classification des textures par les réseaux de neurones convolutifs », mémoire master, université mouloud Mammri tizi-ouzou, 2018/2019.

[23] M.AbderrahmaneAdjila,«Détectiond'activitévocaleutilisantl'apprentissageprofond

[24] »,Mémoiremaster,UniversitédeGhardaïa, 2019/2020.

[25] Apprendreprogrammationcourspython3,(consultéle16/06/2023)disponiblesur

[26] B. Tounsi, "Inférence d'identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte Algérien," Thèse de Magistère, INI ALGER, 2008..

[28] Raschka, S., & Mirjalili, V. ,"Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2" Packt Publishing. ISBN: 978-1-78995-575-0,(2019).